Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 29

A Non-Homogeneous Hidden-State Model on First Order Differences for Automatic Detection of Nucleosome Positions

Pei Fen Kuan* Dana Huebert†

Audrey Gasch‡ Sunduz Keles**

*University of Wisconsin-Madison, kuanp@stat.wisc.edu

†University of Wisconsin-Madison, huebert@wisc.edu

‡University of Wisconsin-Madison, agasch@wisc.edu

**University of Wisconsin-Madison, keles@stat.wisc.edu

A Non-Homogeneous Hidden-State Model on First Order Differences for Automatic Detection of Nucleosome Positions*

Pei Fen Kuan, Dana Huebert, Audrey Gasch, and Sunduz Keles

Abstract

The ability to map individual nucleosomes accurately across genomes enables the study of relationships between dynamic changes in nucleosome positioning/occupancy and gene regulation. However, the highly heterogeneous nature of nucleosome densities across genomes and short linker regions pose challenges in mapping nucleosome positions based on high-throughput microarray data of micrococcal nuclease (MNase) digested DNA. Previous works rely on additional detrending and careful visual examination to detect low-signal nucleosomes, which may exist in a subpopulation of cells. We propose a non-homogeneous hidden-state model based on first order differences of experimental data along genomic coordinates that bypasses the need for local detrending and can automatically detect nucleosome positions of various occupancy levels. Our proposed approach is applicable to both low and high resolution MNase-Chip and MNase-Seq (high throughput sequencing) data, and is able to map nucleosome-linker boundaries accurately. This automated algorithm is also computationally efficient and only requires a simple preprocessing step. We provide several examples illustrating the pitfalls of existing methods, the difficulties of detrending the observed hybridization signals and demonstrate the advantages of utilizing first order differences in detecting nucleosome occupancies via simulations and case studies involving MNase-Chip and MNase-Seq data of nucleosome occupancy in yeast *S. cerevisiae*.

KEYWORDS: nucleosomes, MNase-chip, MNase-Seq, non-homogeneous hidden Markov model, first order differences, smoothing

---

*This research has been supported in part by a PhRMA Foundation Research Starter Grant in Informatics (P.K. and S.K.), the NIH grant HG003747 (P.K. and S.K.), the NSF grant DMS004597 (S.K.), the Morgridge Institute for Research support for Computation and Informatics in Biology and Medicine (P.K), NSF CAREER Award #0447887 (A.P.G and D.J.H.) and an NIGMS grant to the Molecular Biosciences Training Program T32GM007215 (D.J.H.).

1 Introduction

Nucleosomes are the fundamental structural units of chromatin and consist of approximately 146 base pairs of DNA wrapped around a histone octamer (Kornberg and Lorch; 1999; Chakravarthy et al.; 2006). The precise positioning of nucleosomes along the genome has been implicated in the regulation of gene expression (reviewed in Ercan et al. (2004)). Packaging of DNA into nucleosomes may prevent DNA binding proteins from accessing their sites, recruit transcriptional activators or repressors, and bring distant DNA sequences into close proximity to promote transcription (Millar and Grunstein; 2006). In the last couple of years, nucleosome positions have been mapped across the genomes of *S. cerevisiae* (Yuan et al.; 2005; Lee et al.; 2007; Shivaswamy et al.; 2008), *C. elegans* (Johnson et al.; 2006), and humans (Schones et al.; 2008) in various cell types and under a variety of physiological perturbations. These studies have revealed various chromatin remodeling patterns in transcriptional regulation at nucleosome resolution. In particular, Shivaswamy et al. (2008) showed that gene activation in yeast is mainly accompanied by the loss of one or two nucleosomes in the promoter regions, while Lee et al. (2007) illustrated that functionally related genes share similar nucleosome occupancy patterns across their promoters. Nucleosome occupancy can hinder the binding of transcription factors to their consensus motifs, and the fraction of bound motifs vary between nucleosomes and nucleosome free regions (Yuan et al.; 2005). These findings illuminated that identifying locations of individual nucleosomes accurately is essential for studying the effect of dynamic changes in nucleosome occupancy in the control of gene regulation. By having a reliable map of nucleosome occupancy, one can investigate various histone modifications at the nucleosome level to uncover the complex mechanism in transcriptional reprogramming. Similarly, for studies investigating the effect of physiological perturbations on nucleosome positioning, the starting point often involves maps of nucleosome occupancy before and after such perturbations. For example, nucleosome mapping experiments of Schones et al. (2008) in resting and activated human $CD4^+$ T cells revealed specific reorganization patterns of nucleosomes in promoter and enhancer regions of the genome.

Numerous high-throughput experiments have been carried out to map nucleosome occupancy in *S. cerevisiae* via tiling arrays (Liu et al.; 2005; Yuan et al.; 2005; Lee et al.; 2007; Shivaswamy and Iyer; 2008; Kaplan et al.; 2008). More recently, a high resolution whole genome nucleosome map for yeast genome was developed via a high throughput sequencing technology (Albert et al.; 2007; Shivaswamy et al.; 2008). In both platforms, the sample input consists of mono-nucleosomes prepared via micrococcal nuclease (MNase) di-

gestions, which degrades all but the DNA wrapped around histone proteins. Two nucleosomes are connected by linker DNA, which is digested by the enzyme. The digested sample is either sequenced by high-throughput sequencing technologies (MNase-Seq), or competitively hybridized against a control sample using high density tiling arrays in (MNase-Chip). A high percentage of the *S.cerevisiae* genome is known to be occupied by nucleosomes, however there exists substantial variation in nucleosome density across the genome. In particular, relatively higher density of nucleosomes is observed at transcribed regions and lower density is found in intergenic regions (Lee et al.; 2004; Bernstein et al.; 2004; Lee et al.; 2007; Shivaswamy et al.; 2008). Positions of nucleosomes across the whole genome are characterized by a stretch of consecutive probes encompassing approximately 146 base pairs with higher signals than the background. An interesting feature observed in many of the MNase-Chip experiments for mapping nucleosome positions is that the magnitude of log base 2 ratios for regions occupied by nucleosomes exhibit large variability. Specifically, some regions of the genome thought to be occupied by nucloesomes actually show log base 2 ratios below the baseline. Yuan et al. (2005) provided substantial evidence of this problem and referred to this phenomena as unpredictable trends in hybridization. The variability in the magnitudes of nucleosome occupancy is also observable from the high resolution MNase-Chip data of Lee et al. (2007) and MNase-Seq data of Shivaswamy et al. (2008). This trend in hybridization can be attributed to the heterogeneity of nucleosome densities across the whole genome, resulting in both stable and unstable nucleosome occupancies. Unstable or low-signal nucleosomes are nucleosome peaks having low maxima and may correspond to nucleosomes found only in a subpopulation of cells (Yuan et al.; 2005). These low-signal nucleosomes may be the most important and dynamic in regulating transcription by cycling on and off the DNA. We will refer to these as "low-signal nucleosomes" in our subsequent discussion.

Previous work in identifying nucleosome positions in MNase-Chip data include using a hidden Markov model (HMM) (Yuan et al.; 2005) on the observed log base 2 ratios. Yuan et al. (2005) proposed an HMM that takes into account the length of nucleosomal DNA and allows for one emission distribution for each of the nucleosome and linker states, respectively. To account for a global trend, Yuan et al. (2005) applied the HMM to a sliding window of 40 probes and averaged the estimated model parameters and posterior probabilities over all the windows covering a fixed probe to compute the most likely hidden state path. In addition to the nucleosomes identified by the sliding window HMM, they also included additional low-signal nucleosomes obtained by comparing the median intensities of the peak and trough within a window size of 7 probes,

DOI: 10.2202/1544-6115.1454

which was regarded as a detrending procedure to further identify low-signal nucleosomes. We show that applying the proposed detrending procedure to the entire tiling array data is undesirable as it introduces higher noise level and spurious linkers/nucleosomes in simulations and a case study of yeast nucleosome occupancy. Yuan et al. (2005) also provided nucleosome positions hand picked via close visual inspection in order to capture all potential nucleosomes. However, this heuristic approach becomes tedious when mapping nucleosome occupancy in larger genomic regions.

To accommodate the serious drawbacks of existing methods, we propose a fully automated approach which identifies nucleosome occupancy and addresses the length of nucleosomal DNA and the observed trends in hybridization signals. At the core of our methodology is a non-homogeneous hidden-state model (NHSM) tailored for MNase-Chip data measuring nucleosome occupancy. By designing the architecture for the first order (lagged) differences of log base 2 ratios, we bypass the problem of unpredictable trends in the log base 2 ratios. An additional feature of our approach is its applicability to the more recent MNase-Seq data. We illustrate the methodology and benchmark its performance against other available methods in simulations and a case study involving a 20 base pairs resolution yeast MNase-Chip nucleosome occupancy data. The results of these experiments highlight the superiority of our NHSM to other approaches especially in identifying low-signal nucleosomes. We also provide an illustration of its applicability to higher resolution MNase-Chip and MNase-Seq nucleosome occupancy data. Additionally, two consecutive nucleosomes are separated by a linker of variable length. Therefore, a good methodology for mapping nucleosome occupancy should be able to identify nucleosome-linker boundaries accurately. This is usually challenging for the low resolution tiling array design in which a linker is represented by one or two probes (Yuan et al.; 2005; Kaplan et al.; 2008). Our proposed methodology carefully exploits the structure of nucleosomes and accurately maps nucleosome positions and therefore provides a principled framework for studying other epigenetic events that rely on mapping nucleosome occupancy.

2 Motivation

We motivate the idea behind our methodology using the MNase-Chip data from Yuan et al. (2005). We use the normalized median log base 2 ratios of the 8 replicates for illustration. The top panel of Figure 1 shows the nucleosome profile for a region in chromosome 3 in which the nucleosomes identified by Yuan et al. (2005) are marked with black lines (each vertical line represent-

ing a probe), and a stable nucleosome is represented by 6 to 8 probes. It is clear from the plot that the magnitude of log base 2 ratios of a nucleosome region exhibits large variability. Despite having heterogeneous hybridization signals, the plot suggests that a nucleosome is characterized by a peak in the local signal intensity, even if the log base 2 ratio is below the baseline. In other words, a nucleosome occupied region exhibit a "bump" shape irrespective of the actual strength in hybridization signal. In addition, this plot also suggests that using a single distribution for each of the nucleosome and linker/nucleosome depleted regions may fail to distinguish short linkers between stable or well-positioned nucleosomes, (i.e., linkers between well-positioned nucleosomes have comparable hybridization strength to low-signal nucleosomes.)

Given the observed "bump" (or peak with low maxima) characteristic of annotated nucleosomes in the original data, we consider a simple smoothing by replacing the log base 2 ratios of probe i with the average values of probe $i-1$, i and $i+1$. As evident in the middle panel of Figure 1, the "bump" shape is enhanced in the smoothed data which enable easier mapping of the nucleosome positions. The "bumps" also suggest that a nucleosome occupied region is characterized by a series of decreasing positive slopes, followed by slopes of approximately zero in magnitude and then a series of increasing negative slopes. This observation forms the modeling framework of our proposed methodology. The first order differences automatically take care of the trend in hybridization and thereby bring both the low and stable nucleosomes to a comparable level. Once the nucleosome positions are obtained, one can rank the strength of each nucleosome by the average log base 2 ratios of the probes within the nucleosome.

The trend in hybridization and variability in the magnitudes of nucleosome occupancy are not specific only to lower resolution MNase-Chip data of Yuan et al. (2005) but can also be clearly seen in high resolution MNase-Chip data of Lee et al. (2007) and MNase-Seq data of Shivaswamy et al. (2008), as evident in Figures 15 and 17. In addition, we also illustrate that a nucleosome occupied region is characterized by a "bump" shape, which is enhanced upon smoothing in these two data sets (middle and right panels of Figures 15 and 17). The choice of smoothing is presented in Section 3. Since all the different data sets share similar defining characteristics of nucleosome occupancy, for expository purposes, our detailed discussion is mainly focused on the 20 base pairs resolution tiling array data of Yuan et al. (2005). This tiling array design is also used in a recent publication by Kaplan et al. (2008) for studying H3K56 acetylation in yeast. We will demonstrate that our proposed method is applicable and works well in detecting nucleosome positions in both the high resolution tiling arrays and sequencing data in Section 5.2.

DOI: 10.2202/1544-6115.1454

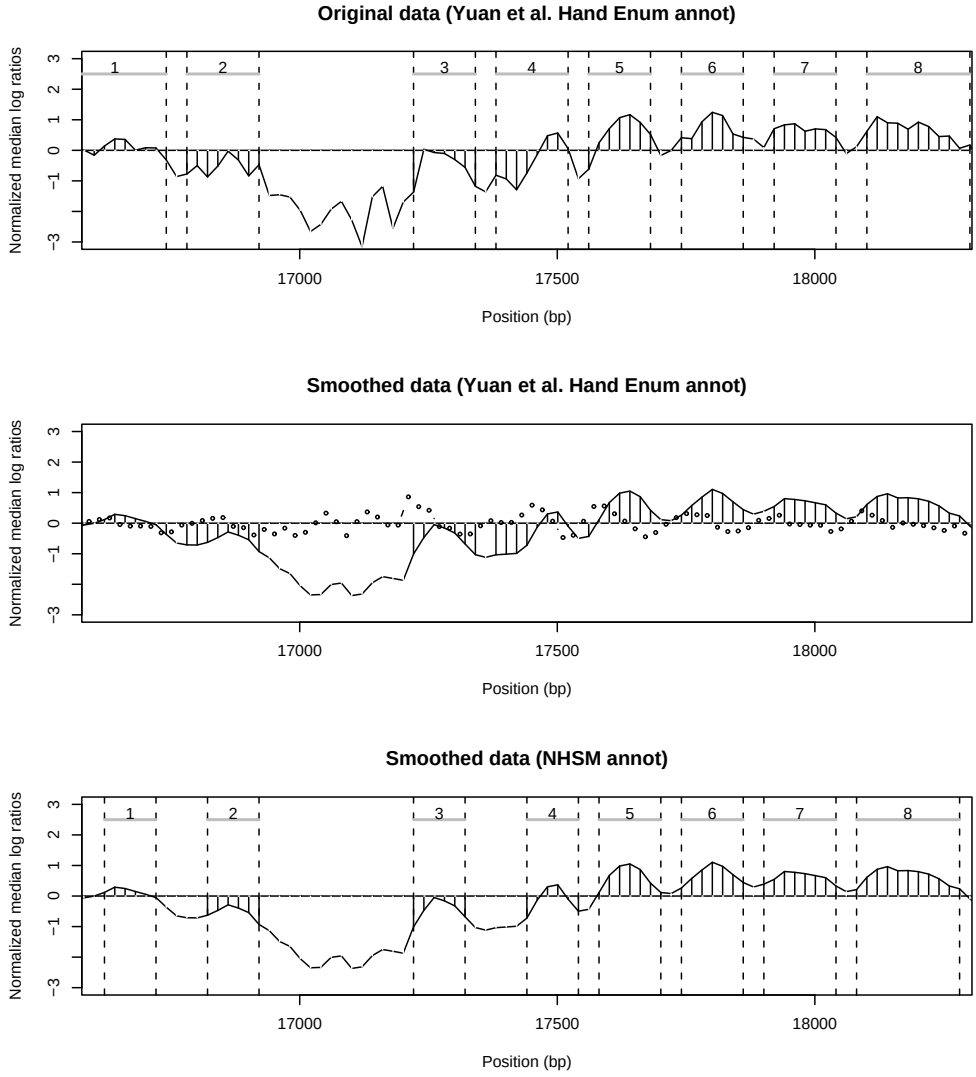


Figure 1: *Typical characteristics of MNase-chip nucleosome occupancy data from Yuan et al. (2005).* Top panel is the original normalized data tiling a region in chromosome 3. The vertical black solid lines represent probes identified as nucleosome state according to "hand picked" annotation in Yuan et al. (2005). The vertical dotted lines are boundaries separating nucleosome-linker states. Gray horizontal lines at y=2.5 are the nucleosomes inferred. Middle panel is the corresponding smoothed data by taking moving averages in a window size of 3 probes and the dots are the first order differences. Bottom panel is based on annotation from our proposed NHSM.

3 Hidden-state model for mapping nucleosome positions

To circumvent the problem of decoding nucleosome occupancy locally to accommodate for the observed local trends as in Yuan et al. (2005), we consider an alternative approach to infer nucleosome positions based on first order (lagged) differences, O_t, which we defined as:

$$\begin{aligned} O_t &= X_{t+k-1} - X_{t-k}, \\ X_t &= S(Y)_t, \end{aligned}$$

where Y_t is the observed log base 2 ratio of probe t and X_t is the corresponding smoothed/denoised data for Y_t, $t = 0, ..., T$. The O_t's are to quantify the slope at midpoint between probe $t-1$ and probe t to capture the "bump" shape of a nucleosome. We define this midpoint location as mid-probe, i.e., mid-probe t corresponds to the midpoint between probe $t-1$ and probe t. Although other choices for defining the slope/gradient can be utilized, we show that the simple first order (lagged) differences is generally sufficient in both the low and high resolution MNase-Chip and MNase-Seq data. For Yuan et al. (2005) data, we let X_t to be a moving average statistic in a window size of $2w + 1$ probes and O_t to be the first order difference ($k = 1$). That is,

$$\begin{aligned} O_t &= X_t - X_{t-1}, \\ X_t &= \sum_{j=t-w}^{t+w} Y_j/(2w+1). \end{aligned}$$

We observe that substituting the log base 2 ratios by the corresponding moving average statistic X_t's reduces the noise in the data and enhances the shape of peaks and troughs, but not at the expense of over smoothing the data as shown in the middle panel of Figure 1. On the other hand, for the high resolution MNase-Chip and MNase-Seq data, the simple moving average can be replaced with a Gaussian kernel smoother. A detailed motivation with some analytical results for using a Gaussian kernel smoothing is given in Appendix A.1. In kernel smoothing, the tuning parameter is the bandwidth h. Large bandwidth implies more smoothing, and vice versa. For Gaussian kernel, h is also the standard deviation. That is,

$$X_t = \frac{\sum_{j=0}^{T} \phi\left(\frac{|G_t - G_j|}{h}\right) Y_j}{\sum_{j=0}^{T} \phi\left(\frac{|G_t - G_j|}{h}\right)},$$

DOI: 10.2202/1544-6115.1454

where $\phi(.)$ is the standard Gaussian probability density function and G_j is the genomic coordinate corresponding to Y_j in base pairs. We propose choosing h based on the size of the nucleosomal DNA. For a Gaussian distribution, 99% of the values are within $\pm 2.5\sigma$, where $\sigma = h$. Therefore, we choose $h = 146/5$ so that the bandwidth spans the size of a nucleosome. The middle and right panels of Figures 15 and 17 illustrate the resulting smoothed log base 2 ratios from Gaussian kernel smoothing with this choice of bandwidth. The Gaussian kernel smoothing is able to denoise the data and enhance the "bump" characteristics of a nucleosome.

As motivated in Section 2, a nucleosome occupied region is characterized by a series of positive followed by negative slopes or O_t's, while the boundaries of nucleosomes-linker regions are characterized by steeper slopes. Detecting jumps in O_t's via segmentation is a potential approach to map nucleosome occupancy but traditional segmentation approaches do not incorporate the length of nucleosomal DNA. In addition, since the data is obtained from tiling arrays, spatial correlations among observations of nearby probes are expected. To account for the length of nucleosomal DNA and the correlation structure, we propose a non-homogeneous hidden-state model (NHSM) based on first order differences O_t's. Next, we give a detailed characterization of the NHSM architecture.

Consider the state transitions given in Figure 2(a) where N_i's represent the nucleosome region states, L_i's represent linker or nucleosome depleted region state and B_i's represent nucleosome-linker boundaries. The self transitions of N_1 and N_3 is to account for less stable nucleosomes which span a larger region than well-positioned nucleosomes, termed "fuzzy" nucleosomes by Yuan et al. (2005). Recently, Valouev et al. (2008) also provided ample evidence for the existence of fuzzy nucleosomes in the *C. elegans* genome using high-throughput sequencing data. They attributed this to the lack of constraints in absolute positioning for some fraction of the nucleosomes across the *C. elegans* genome. Therefore, we introduce state duration $d(i)$ to capture the length of nucleosomal DNA explicitly. Assume that a well-positioned nucleosome (146 base pairs) is characterized by p probes, or equivalently $p-1$ first order differences. We require

$$\sum_{i \in \{N_{2a}, N_{2b}, N_{2c}\}} d(i) + 2 \;=\; p-1, \quad 0 \le d(N_{2a}), d(N_{2b}) \le p-3,$$

since at least one probe is from N_1 and one is from N_3 out of $p-1$ probes representing a nucleosome.

In most cases, the "bump" shape of a nucleosome on tiling arrays is symmetrical, which implies that $d(N_{2a}) = d(N_{2c})$. Moreover, given the state du-

ration constraint, the state transitions can be further simplified as in Figure 2(b) by tying states N_{2a}, N_{2b} and N_{2c} as N_2 with a trinomial duration density:

$$p_{N_2}(d_1, d_2, d_3) = \frac{(p-3)!}{d_1!d_2!d_3!} p_1^{d_1} p_2^{d_2} p_3^{d_3},$$

where $p_1 + p_2 + p_3 = 1$ and $d_1 + d_2 + d_3 = p - 3$.

Let $b_i(O_t)$ denote the emission distribution for observed value at mid-probe $t = 1, ..., T$ given unknown state $i \in \{N_i, L_i, B_i\}$. We model $b_i(O_t)$ with Gaussian distributions,

$$\begin{aligned}
b_{B_N}(O_t) &\sim N(\mu_1, \sigma^2_{B_N}), & b_{N_1}(O_t) &\sim N(\mu_2, \sigma^2_{N_1}),\\
b_{N_3}(O_t) &\sim N(-\mu_2, \sigma^2_{N_5}), & b_{B_L}(O_t) &\sim N(-\mu_1, \sigma^2_{B_L}),\\
b_{L_1}(O_t) &\sim N(-\mu_2, \sigma^2_{L_1}), & b_{L_2}(O_t) &\sim N(0, \sigma^2_{L_2}),\\
b_{L_3}(O_t) &\sim N(\mu_2, \sigma^2_{L_3}), & b_{N_2}(O_{t:t+p-4}) &\sim N(\tilde{\mu}, \Sigma),
\end{aligned}$$

where

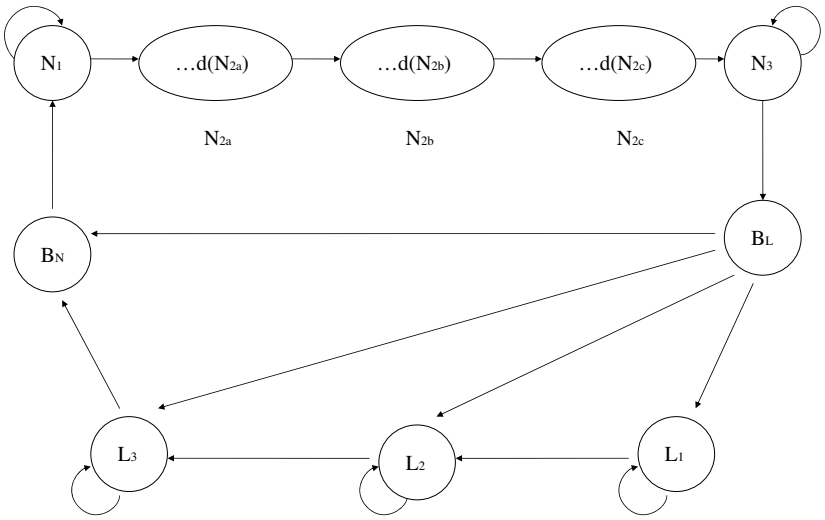
$$\begin{aligned}
\tilde{\mu} &= (\underbrace{\mu_2, ..., \mu_2}_{d_1}, \underbrace{0, ..., 0}_{d_2}, \underbrace{-\mu_2, ..., -\mu_2}_{p-3-d_1-d_2}),\\
\Sigma &= \mathrm{diag}(\underbrace{\sigma^2_{N_{2a}}, ..., \sigma^2_{N_{2a}}}_{d_1}, \underbrace{\sigma^2_{N_{2b}}, ..., \sigma^2_{N_{2b}}}_{d_2}, \underbrace{\sigma^2_{N_{2c}}, ..., \sigma^2_{N_{2c}}}_{p-3-d_1-d_2}),
\end{aligned}$$

and $0 < \mu_2 < \mu_1$. The constraint on the mean of emission distributions is to ensure the series of decreasing positive slopes, zero slopes and followed by increasing negative slopes which characterize the "bump" shape of a nucleosome. In the case of symmetric "bump" shape, the duration density for N_2 reduces to univariate density $p(d_1)$ and

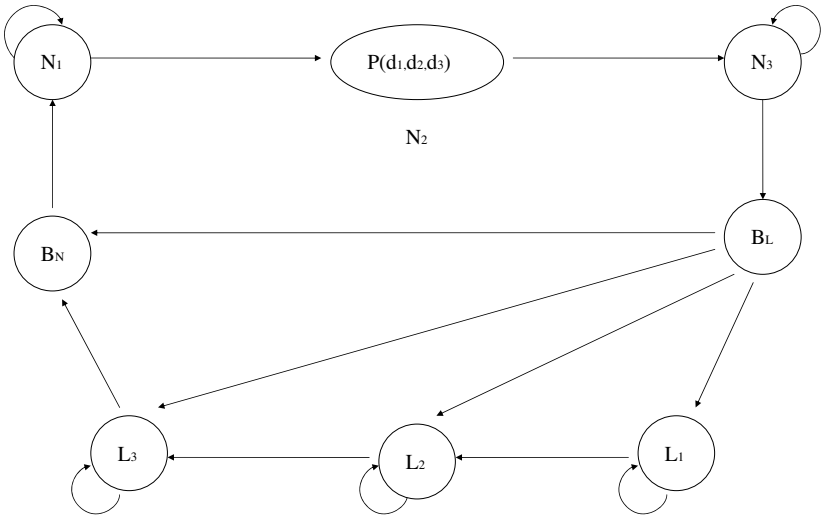
$$\tilde{\mu} = (\underbrace{\mu_2, ..., \mu_2}_{d_1}, \underbrace{0, ..., 0}_{p-3-2d_1}, \underbrace{-\mu_2, ..., -\mu_2}_{d_1}).$$

The discrete duration density assumption implies that the proposed duration in the hidden states is equivalent to a hidden-state model with a larger hidden state space. We can recast the state transition in Figure 2(a) as Figure 3 which have the same complexity by considering all possible uni-directional paths transiting from N_1 and incorporating the constraint $\sum_{i \in \{N_{2a}, N_{2b}, N_{2c}\}} d(i) + 2 = p - 1$. We can equivalently let $b_{N_{2a}}(O_t) \sim N(\mu_2, \sigma^2_{N_{2a}})$, $b_{N_{2b}}(O_t) \sim N(0, \sigma^2_{N_{2b}})$ and $b_{N_{2c}}(O_t) \sim N(-\mu_2, \sigma^2_{N_{2c}})$. In scenarios where we have high resolution experiments for mapping nucleosome occupancy such as the 4 base pairs resolution MNase-chip data of Lee et al. (2007) or 1 base pair resolution

DOI: 10.2202/1544-6115.1454



(a)



(b)

Figure 2: *State transition representation in NHSM.* N_i represents nucleosome states, L_i represents linker states, B_N and B_L represent linker-nucleosome and nucleosome-linker boundaries, respectively.

MNase-seq data of Shivaswamy et al. (2008), the "bump" shape of nucleosome is relatively well characterized by a few positive slopes, followed by a plateau and a few negative slopes. In such cases, we can reduce the range of d_1 by removing some uni-directional paths in Figure 3 and thereby simplify the structure of the hidden state transitions.

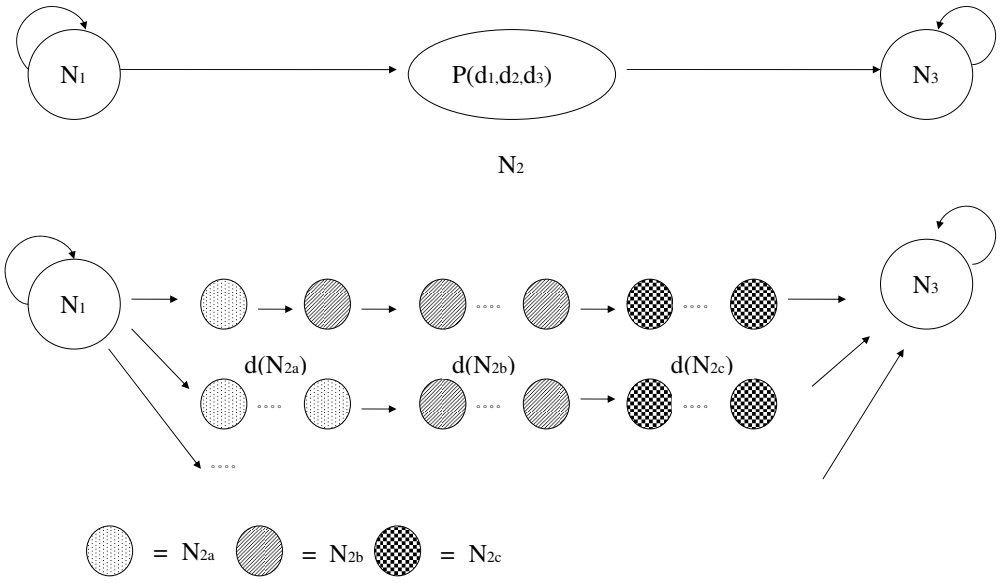


Figure 3: *State transition representation in NHSM.* An equivalent representation of the discrete duration density in the hidden states of Figure 2(a).

Let Q_t denote the hidden state for mid-probe t. Note that if $Q_t = B_N$, this indicates that probe $t-1$ is in linker region and probe t is in nucleosome region. Since high log base 2 ratios represent regions that are more likely to be occupied by nucleosomes and vice versa for low log base 2 ratios, we model the hidden state transitions as a function of observed log base 2 ratios X_t. This framework implies that while the nucleosome/linker hidden state Q_t dictates the observed first order differences (a function of X_t), there is another function of X_t which in turn influences the hidden states and gives rise to a feedback structure. This idea is adapted from Zucchini et al. (2008) who proposed a

DOI: 10.2202/1544-6115.1454

mechanism to allow for such feedback. We refer the readers to Zucchini et al. (2008) for an excellent motivation of this framework in the context of animal behavior. We define $Z_t = O_t + Z_{t-1}$ for $t = 1, ..., T$ and $Z_0 = X_0$. In the case where $O_t = X_t - X_{t-1}$, we have $Z_t = X_t$. This feedback structure is best understood by considering the following graphical model representation (Figure 4).

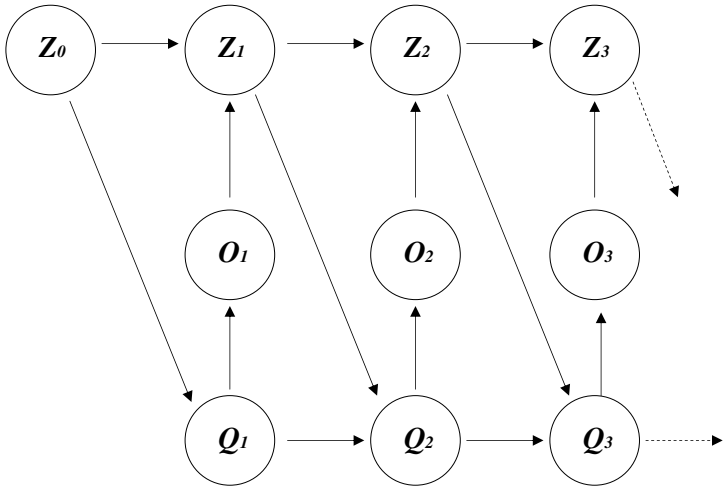


Figure 4: *Graphical model representation of the feedback structure.* The directionality of the edges dictates the dependence structure. This model implies that the transition from Q_t to Q_{t+1} depends on Z_t.

Let $a_{i,j}(z) = P(Q_{t+1} = j \mid Q_t = i, Z_t = z)$ be the transition probabilities from state i to j between mid-probe t and mid-probe $t+1$ given Z_t. Also, write $O^{(t)} = (O_1, ..., O_t)$. Two assumptions arising from Figure 4 are:

$$P(Q_{t+1}|Q^{(t)}, Z_0, O^{(t)}) = P(Q_{t+1}|Q_t, Z_t) \text{ for } t = 1, ..., T-1. \quad (1)$$

$$P(O_t|Q^{(t)}, Z_0, O^{(t-1)}) = P(O_t|Q_t) \text{ for } t = 1, ..., T. \quad (2)$$

To avoid overparametrization, only transitions $a_{B_L,\bullet}(Z_t)$, $a_{L_3,\bullet}(Z_t)$ and $a_{N_3,\bullet}(Z_t)$ are functions of Z_t's. Other transition probabilities are assumed to be time homogeneous. We employ a logistic regression model to parametrize the hidden transitions for B_L, L_3 and N_3:

$$a_{i,j}(Z_t) = \frac{\exp(\gamma_{i,j} + \beta_j Z_t)}{\sum_{k=1}^{N} \exp(\gamma_{i,k} + \beta_k Z_t)}.$$

In cases where the data has been median centered at zero, we observe that a simpler version of the non-homogeneous transition probabilities for these three hidden states performs well (see case study). That is, we consider

$$a_{i,j}(Z_t) \;=\; \begin{cases} a^n_{i,j}, & \text{if } Z_t < 0, \\ a^p_{i,j}, & \text{if } Z_t \geq 0. \end{cases}$$

For instance, we can let $a_{L_3,B_N}(Z_t) = I(Z_t \geq 0) + a^n_{L_3,B_N} I(Z_t < 0)$ to impose transition into nucleosome states when $Z_t \geq 0$. Details of model fitting are given in Appendix A.2.

4 Simulation studies

Yuan et al. (2005) attributed the heterogeneous nucleosome density to unpredictable trends in hybridization data. They applied the HMM to a sliding window of 40 consecutive probes to address this issue. Hidden states decoding via the Viterbi algorithm was based on average values of the model parameters and posterior probabilities of all windows containing a fixed probe. We referred to this method as sliding window HMM (SHMM). SHMM is computationally intensive and requires one to select the window size, which depends on the trend in hybridization. Yuan et al. (2005) also proposed detrending the data by comparing the magnitude of peak and trough locally to capture low-signal nucleosomes. In particular, for each probe, they considered a window size of 7 probes ($\sim$ size of a nucleosome) centered at the probe and replaced the observed log base 2 ratio by the difference between the median of log base 2 ratios within the window and the minimum log base 2 ratio of the two probes adjacent to this window. They observed that the trend was effectively eliminated using this procedure. We referred to this method as HMMD (detrending followed by usual HMM to infer nucleosome/linker states).

4.1 Simulation I: Hidden Markov model with trend line

In the first simulation, we generated the data using the HMM hidden states architecture in Figure 5(a) (or Figure S1E of Yuan et al. (2005)), in which well-positioned nucleosomes were represented by 6 to 8 probes (N1-N8) and delocalized nucleosomes (D1-D9) covered at least 9 probes. Nucleosome regions were expected to have high log base 2 ratios whereas linker regions had lower values. The hidden state transitions in Yuan et al. (2005) allowed for linker regions (L) to have variable length. Conditioned on the hidden states,

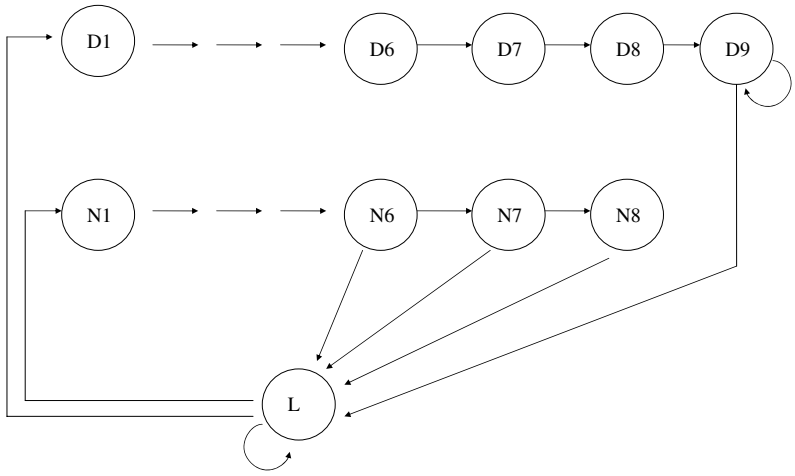
DOI: 10.2202/1544-6115.1454

the observed log base 2 ratios were generated from Gaussian distributions, with mean 0.7, standard deviation (s.d.) 0.2 for nucleosome states and mean -0.7, s.d. 0.3 for linker state. We illustrated that although we were simulating the observed log base 2 ratios, and not the first order differences, our proposed NHSM was able to map nucleosome positions accurately.

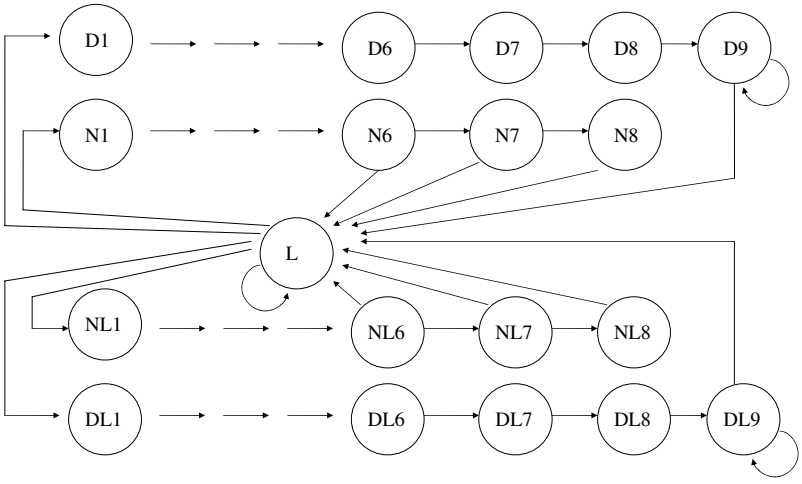
To simulate heterogeneous nucleosome densities, we added a trend line to the simulated data following Yuan et al. (2005). Figure 1 suggests that the underlying trend line in the observed data resembles a curve. Therefore, instead of adding a linear trend line as in Yuan et al. (2005), we let the trend be a sinusoidal curve so that the synthetic data resembles the observed data to a larger extent (Figure 6 top right panel). The bottom left panel of Figure 6 plots the detrended data obtained by comparing peak to trough in a window size of 7 described above. Although this procedure was able to remove the trend in hybridization, it introduced artificial linkers within delocalized nucleosomes and spurious "bumps" within nucleosome depleted/long linker regions and resulting in data with higher noise level. This suggests that applying the same detrending procedure to the whole data is not desirable. On the other hand, a simple smoothing of the synthetic data preserved the "bump" shape that characterizes a nucleosome (Figure 6 bottom right panel). We considered sinusoidal curves with different periodicity (Figure 7) in this simulation study.

4.2 Simulation II: Hidden Markov model with mixture emission distributions

Although adding a trend line results in sythetic data that resembled the actual observed data, it may not be the most realistic model to describe the heterogeneity of nucleosome densities. We considered a more realistic simulation setup to generate nucleosomes with various occupancy levels by using mixture emission distributions for the hidden states. We enlarged the hidden state transitions (Figure 5(b)) by introducing low and high (stable) nucleosome states. The stable nucleosomes (N1-N8, D1-D9) were generated from a Gaussian distribution with mean 0.7 and s.d. 0.2. Low-signal nucleosomes (NL1-NL8, DL1-DL8) were generated from a Gaussian distribution with mean 0.1 and s.d. 0.3 and the linker state was generated from a mixture of 3 Gaussian distributions with means -0.3, -0.5, -0.7 and constant s.d. 0.3 with equal mixing proportion. An example of simulated data is shown in Figure 8. The middle panel again shows that detrending introduces a higher noise level to the original data.



(a)



(b)

Figure 5: *HMM architecture in Yuan et al. (2005).* D1-D9 represent delocalized high nucleosomes, N1-N8 represent well-positioned high nucleosomes, DL1-DL9 represent delocalized low-signal nucleosomes, NL1-NL8 represent well-positioned low-signal nucleosomes and L represents a linker probe.

DOI: 10.2202/1544-6115.1454

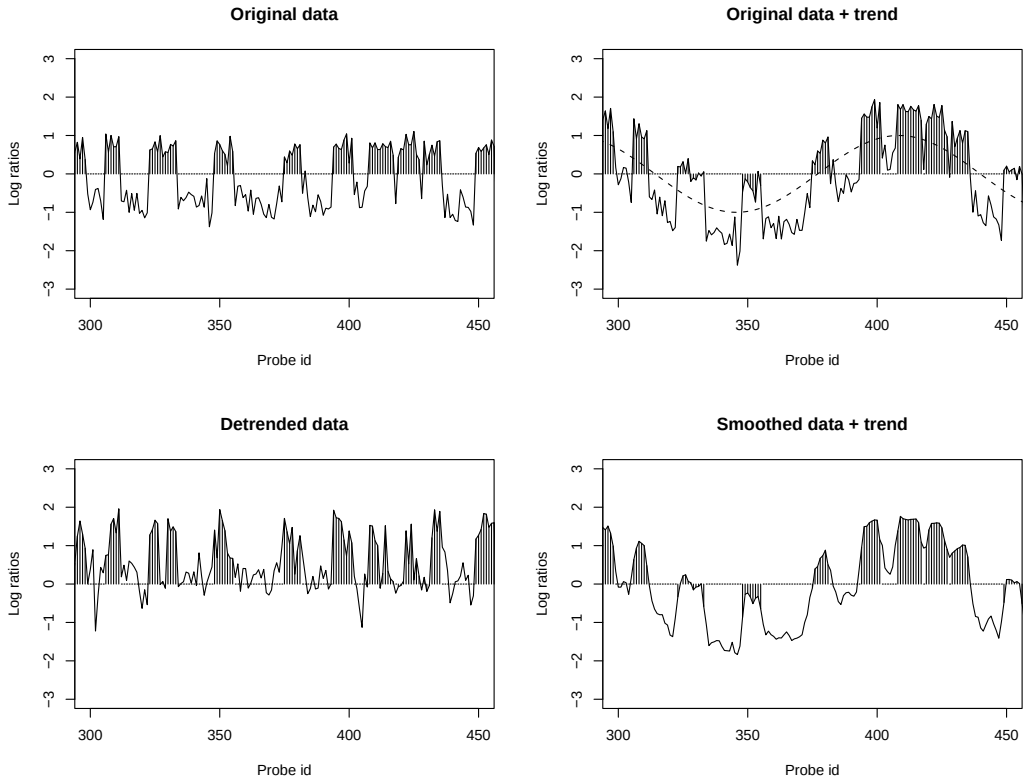


Figure 6: *An example of simulated data from Simulation I.* The dotted line in the top right panel is the trend line. Bottom left panel is the data detrended by comparing peak and trough within a window size of 7 probes. Bottom right panel is the smoothed data. Black vertical lines represent true nucleosome probes.

We simulated observations for 1000 probes according to a tiling design of 50-mer probes overlapped by 30 base pairs covering a 20030 base pair region. In both simulations, we decoded the hidden states using the usual HMM with two emission distributions, one for linker and one for nucleosomes (without differentiating fuzzy/well-positioned, low/high), SHMM, HMMD (detrend first, then apply usual HMM) and our proposed NHSM (on first order differences). The most probable path for each method was decoded via the Viterbi algorithm (Appendix A.2.3).

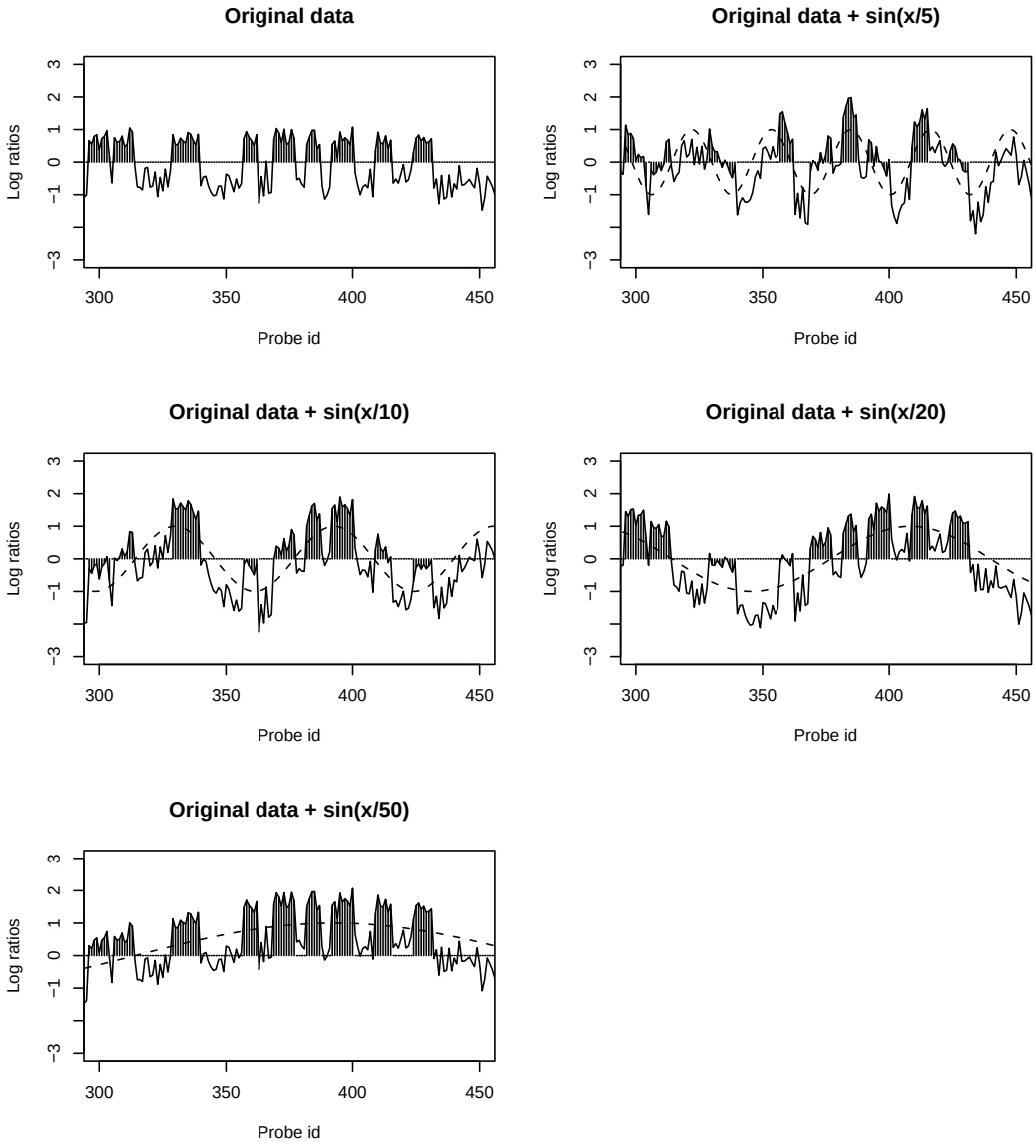


Figure 7: *Trend lines in the simulated data.* The periodicity of the sinusoidal trend lines is varied in each simulation scenario.

DOI: 10.2202/1544-6115.1454

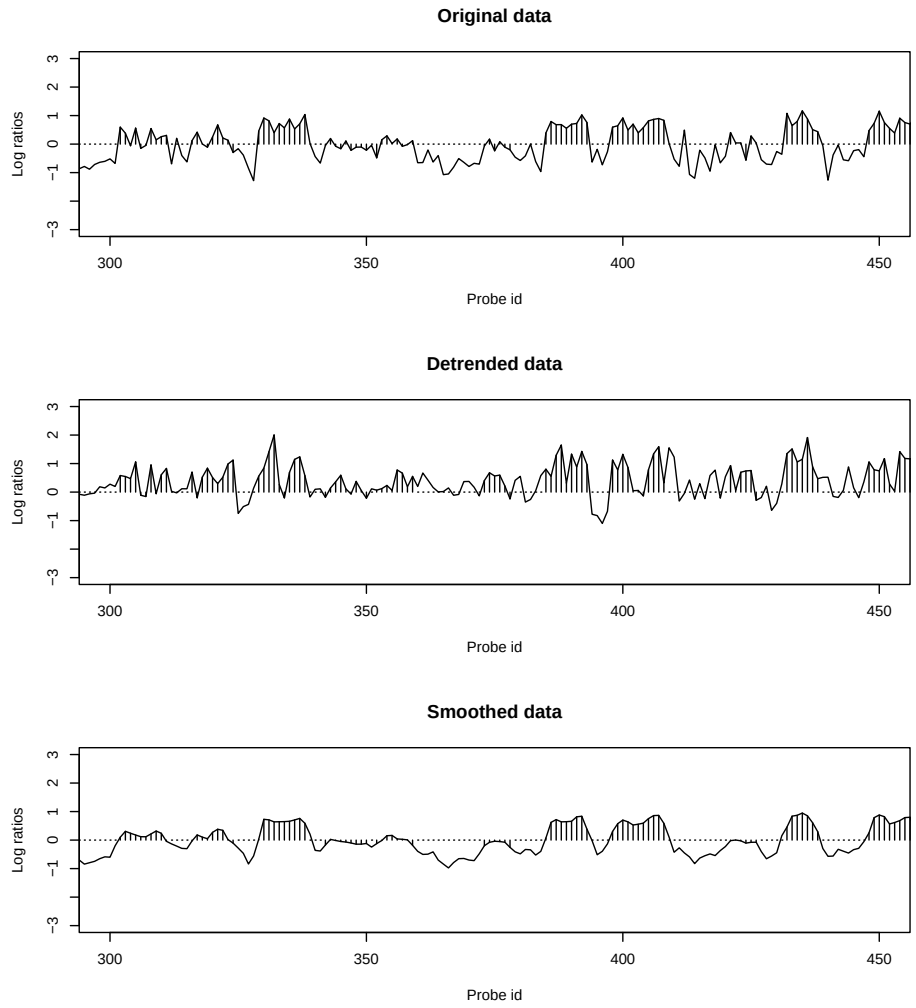


Figure 8: *An example of simulated data from Simulation II.* Middle panel is the data detrended by comparing peak and trough within a window size of 7 probes. Bottom panel is the smoothed data. Black vertical lines represent nucleosome probes.

4.3 Results

We compared the performance of each method via the area under a receiver operating characteristic (AUROC) curve, by varying the posterior probabilities of declaring a probe to be in a nucleosome (well positioned and delocalized) state. In addition, we also evaluated the sensitivity and specificity at probe level of the most probably path for each method. The results, averaged over 50 simulated data sets of 1000 probes, are summarized in Table 1.

In both simulations, NHSM has a consistent result and outperforms other methods in both the sensitivity/specificity at the 0.5 posterior probability threshold and AUROC, since its main assumption is the "bump" shape that characterizes a nucleosome and this characteristic is preserved irrespective of the underlying trends in hybridization (Simulation I). HMM consistently tends to declare fewer nucleosomes, resulting in lower sensitivity. On the other hand, in cases where the trend line has larger periodicity, comparing the magnitudes of peaks and troughs is able to remove the trend effect and improves the performance of HMMD, although it is still worse than NHSM. The superior performance of SHMM in Simulation I with larger periodicity is not surprising. When the periodicity is large, the simulated data in each segment consisting of 40 probes is very close the the original hidden Markov model generator with scaled mean in the emission distributions, and therefore fitting a usual HMM to each segment in SHMM agrees with the underlying data generator. However, when the trend line oscillates more frequently (Simulation I) or is unpredictable (Simulation II), the performance of SHMM decreases rapidly. This indicates that the sliding window size in SHMM depends heavily on the trend in hybridization. In the actual data analysis, it is hard to calibrate the window size since the exact trend is unknown, and a reasonable number of probes within the window size is required for obtaining reliable parameter estimates in an HMM fit.

5 Case studies

5.1 Mapping nucleosome occupancy in MNase-chip data

We illustrated our proposed NHSM on the normalized median log base 2 ratios of the 8 replicates from Yuan et al. (2005). The data was generated from microarrays which consist of 50-mer oligonucleotides probes tiled at 20 base pairs resolution, covering approximately half megabase of the yeast genome. A moving average in a window size of 3 probes was first applied across the whole data as the smoothing step. A well positioned nucleosome ($\sim$146 base pairs) is represented by at least 6 probes (Yuan et al.; 2005), which implies that $0 \leq d(N_{2a}), d(N_{2b}), d(N_{2c}) \leq 3$. We also assumed that $d(N_{2a}) = d(N_{2c})$. Therefore, the structure of state transitions in the hidden states is simplified and given in Figure 9. For this case study, we considered the simpler non-

DOI: 10.2202/1544-6115.1454

Trend	Method	Sensitivity	Specificity	AUROC
$\sin(x/5)$	HMM	0.527 ± 0.095	0.937 ± 0.094	0.718 ± 0.056
	SHMM	0.671 ± 0.042	0.783 ± 0.039	0.821 ± 0.021
	HMMD	0.596 ± 0.064	0.903 ± 0.045	0.786 ± 0.037
	NHSM	0.874 ± 0.051	0.873 ± 0.031	0.962 ± 0.010
$\sin(x/10)$	HMM	0.501 ± 0.065	0.969 ± 0.081	0.727 ± 0.048
	SHMM	0.721 ± 0.048	0.886 ± 0.054	0.870 ± 0.022
	HMMD	0.788 ± 0.043	0.898 ± 0.028	0.949 ± 0.012
	NHSM	0.956 ± 0.028	0.927 ± 0.032	0.986 ± 0.005
$\sin(x/20)$	HMM	0.542 ± 0.100	0.909 ± 0.144	0.717 ± 0.077
	SHMM	0.989 ± 0.006	0.992 ± 0.012	0.997 ± 0.004
	HMMD	0.814 ± 0.032	0.917 ± 0.015	0.917 ± 0.015
	NHSM	0.966 ± 0.025	0.922 ± 0.028	0.988 ± 0.006
$\sin(x/50)$	HMM	0.542 ± 0.086	0.959 ± 0.112	0.738 ± 0.064
	SHMM	0.998 ± 0.003	0.998 ± 0.003	0.999 ± 0.001
	HMMD	0.817 ± 0.031	0.899 ± 0.024	0.963 ± 0.007
	NHSM	0.969 ± 0.025	0.943 ± 0.023	0.988 ± 0.005
mixture emission	HMM	0.564 ± 0.131	0.996 ± 0.010	0.731 ± 0.044
	SHMM	0.834 ± 0.036	0.967 ± 0.022	0.969 ± 0.007
	HMMD	0.571 ± 0.107	0.902 ± 0.080	0.751 ± 0.096
	NHSM	0.928 ± 0.055	0.967 ± 0.016	0.987 ± 0.005

Table 1: *Mean sensitivity, mean specificity and AUROC from the 50 simulations with the corresponding standard errors for each method.* Sensitivity and specificity calculations are based on the most probably path decoding in each method. AUROC illustrates the overall performance across the range of all posterior probabilities cut-offs.

parametric transition probabilities for B_L, L_3 and N_3:

$$\begin{aligned} a_{N_3,B_L}(Z_t) &= \begin{cases} 1, & \text{if } Z_t < 0, \\ a^p_{N_3,B_L}, & \text{if } Z_t \geq 0, \end{cases} \\ a_{B_L,B_N}(Z_t) &= \begin{cases} a^n_{B_L,B_N}, & \text{if } Z_t < 0, \\ 1, & \text{if } Z_t \geq 0, \end{cases} \\ a_{L_3,B_N}(Z_t) &= \begin{cases} a^n_{L_3,B_N}, & \text{if } Z_t < 0, \\ 1, & \text{if } Z_t \geq 0. \end{cases} \end{aligned}$$

where $Z_t = O_t + Z_{t-1}$. Since $O_t = X_t - X_{t-1}$ in this case study, we have $Z_t = X_t$.

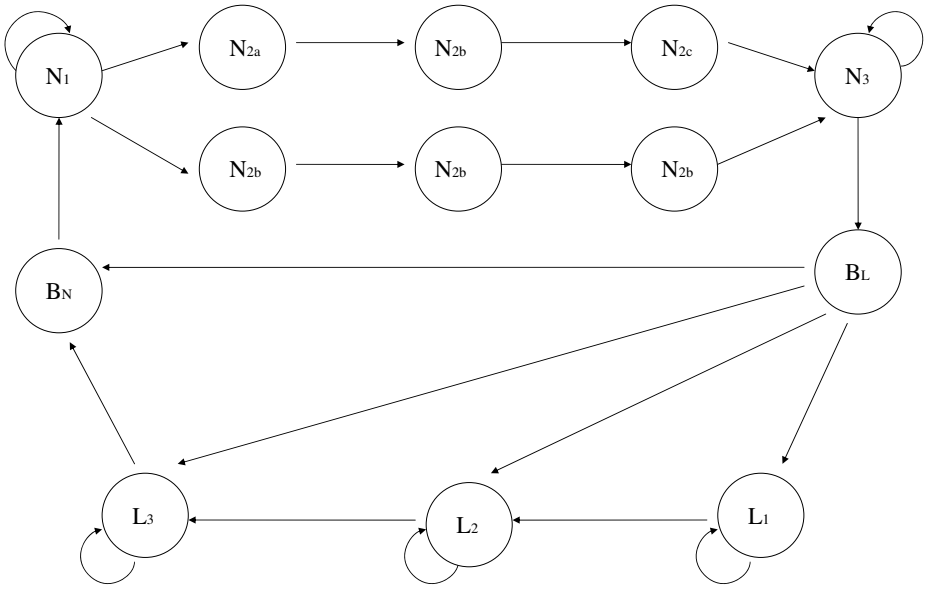


Figure 9: *Simplified state transition representation in NHSM for MNase-chip data of Yuan et al. (2005).* We assume that $d(N_{2a}) = d(N_{2c})$.

This transition structure implies that if the current state is in a linker region, a positive log base 2 ratio observed in the immediate probe imposes transition into a nucleosome state. Similarly, if the current state is in N_3 nucleosome state, a negative log base 2 ratio observed in the immediate probe imposes transition into a linker state. This transition structure appears to be sufficient and works well on the data. We first illustrated that our proposed NHSM is able to detect low-signal nucleosomes in the *HIS3* promoter region as shown in Figure 10. The horizontal black line between positions 721871 and 721971 is the low-signal nucleosome annotated in Figure 1B of Yuan et al. (2005) which was only identified by further detrending the data for low-signal nucleosomes. In particular, Yuan et al. (2005) first applied SHMM to decode nucleosome positions, and further included additional nucleosomes identified exclusively only by HMMD, which they labeled as low-signal nucleosomes. Although Yuan et al. (2005) used the SHMM annotation as baseline nucleosomal set and added low-signal nucleosomes from HMMD to this set, this procedure could be problematic if the boundaries of the low-signal nucleo-

DOI: 10.2202/1544-6115.1454

somes cut across the boundaries of flanking nucleosomes identified in SHMM. In general, combining the two sets of annotation requires careful curation in determining the boundaries of nucleosomes and linkers since the two annotations do not coincide precisely. This low-signal nucleosome was also identified by others according to Yuan et al. (2005) and in the high resolution MNase-Chip experiment of Lee et al. (2007) and MNase-Seq experiment of Shivaswamy et al. (2008), therefore it is not likely to be an artifact of hybridization. Our proposed NHSM is able to map this low-signal nucleosome automatically and accurately without any additional detrending. We provided an example of a low-signal nucleosome that was still missed by further detrending in Yuan et al. (2005) in Figure 11. This low-signal nucleosome was also annotated in high resolution data of Lee et al. (2007) and Shivaswamy et al. (2008) and this provides evidence against it being a hybridization artifact. We also showed that the duration constraint in nucleosome states in our NHSM architecture is able to distinguish real "bumps" which characterize a nucleosome from spurious small "bumps" at positions 103400 (between nucleosomes 1 and 2) and 104400 (between nucleosomes 6 and 7) in the top panels of Figure 12. The problem with detrending the data by comparing peak and trough within a window size of 7 probes is also visible in this region. As evident in the bottom left panel of Figure 12, detrending introduced more noise to the original data and diminished the distinction between linker and nucleosomes.

To compare the performance of the different methods, we used the "hand picked" annotation in Yuan et al. (2005) as the gold standard. Hand picked annotation was based on careful visual inspection (Yuan et al.; 2005), and thus formed a reliable nucleosome map for this tiling array data. However, it is inevitable that there may still exist some uncertainties in mapping nucleosome-linker boundaries even by careful visual inspection as shown in Figure 13. To account for the one/two probes boundary uncertainties in the "hand picked" annotation, we allowed for one probe margin in defining sensitivity and specificity. That is, suppose that the underlying state for probe i is nucleosomal based on "hand picked" annotation. We declare this probe to be correctly inferred if either one of the probes $i-1$, i or $i+1$ is annotated as nucleosomal probe for each of the method under comparison. To measure the sensitivity of our proposed method in detecting low-signal nucleosomes, we considered two possible sets of true positives. The first set was defined by using probes annotated as nucleosomes (both low and high signals) in the "hand picked" annotation. The second set was defined by using probes categorized as low-signal nucleosomes by "hand picked" annotation according to Yuan et al. (2005) (that is corresponding to score 0.25 and 0.5 in Yuan et al. (2005)).

Table 2 summarizes the sensitivity and specificity for these methods using "hand picked" annotation as the gold standard. "Sensitivity(both)" was obtained using all annotated nucleosomes as true positives while "Sensitivity(low)" was obtained using annotated low-signal nucleosomes as true positives. SHMM misses a very large fraction of the low-signal nucleosomes, and thereby has extremely poor sensitivity. HMM has a higher sensitivity than SHMM, but a much lower specificity. The methods are comparable in terms of their specificities, except for HMM. The sensitivity analysis illustrates that the proposed NHSM based on first order differences is able to bypass the need for local detrending and automatically map nucleosome positions accurately. HMMD is the worst among all, which again illustrates that detrending the data is a difficult procedure and could potentially distort the signals in the observed data.

We also compared the performance of our proposed NHSM, HMM, HMMD and SHMM (from Yuan et al. (2005)) via ROC curves, by varying the posterior

Method	Sensitivity(both)	Sensitivity(low)	Specificity
HMM	0.905	0.547	0.784
SHMM	0.849	0.231	0.965
HMMD	0.654	0.519	0.753
NHSM	0.937	0.909	0.934

Table 2: *Sensitivity/specificity for the case study.* Sensitivity and specificity are computed by treating the "hand picked" annotation of Yuan et al. (2005) as the gold standard.

probabilities of declaring a probe to be in a nucleosome (well positioned and delocalized) state using the low-signal nucleosomes as true positive set. The results are shown in Figure 14, which demonstrates that the proposed NHSM based on first order differences performs better than all the other methods. Specifically, NHSM has an AUROC statistic of 0.889 whereas the best of the other methods has only a value of 0.808.

The above benchmarking study relied on using the "hand picked" annotation of Yuan et al. (2005) as the gold standard. We also provide an additional benchmarking experiment in Table 3 by annotating the probes in Yuan et al. (2005) based on the nucleosome calls from the higher resolution experiment of Lee et al. (2007) which used a 4 base pairs resolution tiling array. Comparisons of different methods using this annotation as the gold standard again indicate that NHSM uniformly outperforms other methods by identifying a larger percentage of low-signal nucleosomes.

DOI: 10.2202/1544-6115.1454

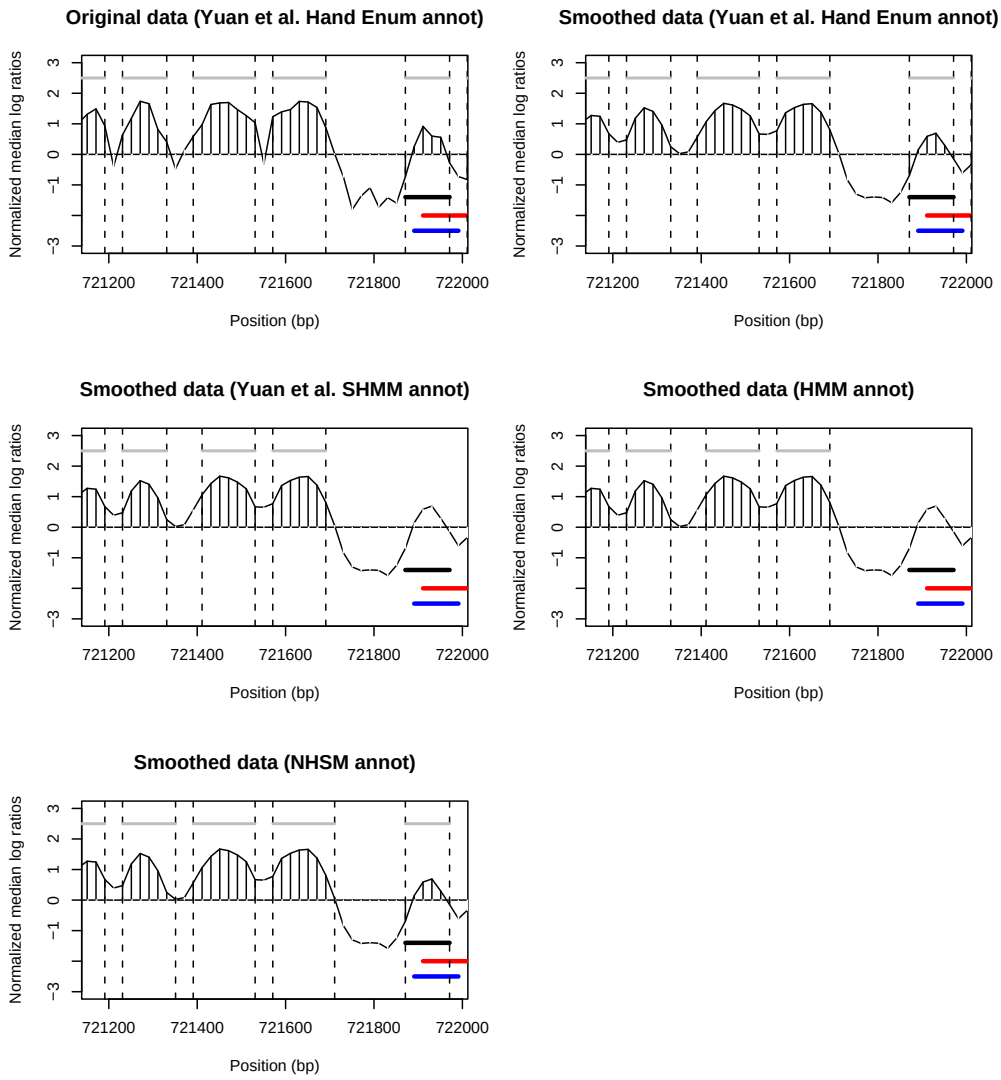


Figure 10: *Nucleosome occupancy in HIS3 promoter.* Top left panel is the original normalized data tiling *HIS3* promoter region and using annotation based on"hand picked" nucleosomes in Yuan et al. (2005). Top right panel is similar to top left panel except that we plot the corresponding smoothed data. Middle left and right panels are based on SHMM annotation in Yuan et al. (2005) and ordinary HMM annotation, respectively. Bottom left panel is based on annotation from our proposed NHSM. Black horizontal line between positions 721871 and 721971 in each panel is the low nucleosome identified by Yuan et al. (2005) after further detrending. Red and blue horizontal lines are the nucleosome regions identified independently by Lee et al. (2007) and Shivaswamy et al. (2008), respectively.

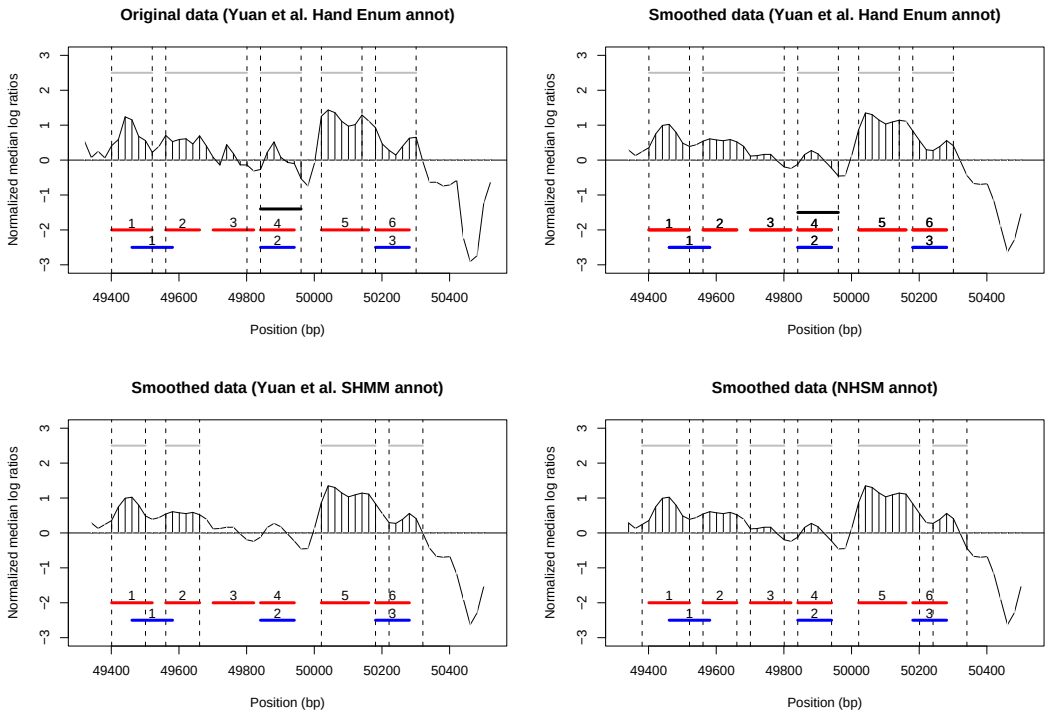


Figure 11: *An example of "hand picked" low-signal nucleosome for a region in chromosome 3.* Black horizontal line between positions 49841 and 49961 is an example of "hand picked" low-signal nucleosome by Yuan et al. (2005). Red and blue horizontal lines are the nucleosome regions identified independently by Lee et al. (2007) and Shivaswamy et al. (2008). The additional detrending by Yuan et al. (2005) after SHMM decoding still misses some of the low-signal nucleosomes, but NHSM is able to capture them.

Method	Sensitivity	Specificity
HMM	0.893	0.653
SHMM	0.815	0.796
HMMD	0.619	0.657
NHSM	0.914	0.784

Table 3: *Sensitivity/specificity for the case study using annotations from Lee et al. (2007).* Sensitivity and specificity are computed by treating the annotation of Lee et al. (2007) as the gold standard.

DOI: 10.2202/1544-6115.1454

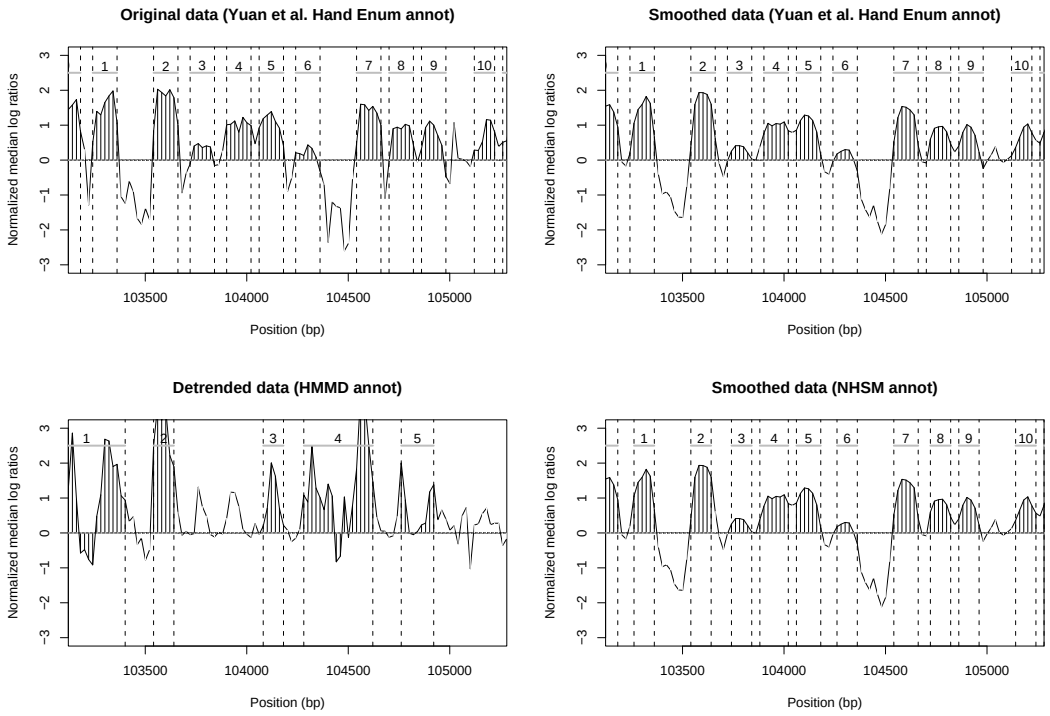


Figure 12: *Nucleosome occupancy for a region in chromosome 3 in Yuan et al. (2005).* Top panels are based on "hand picked" annotation. Bottom left panel is the detrended data by comparing peak and trough within a window size of 7 probes. Bottom right panel is based on annotation from our proposed model. The spurious "bumps" at positions 103400 (between nucleosomes 1 and 2) and 104400 (between nucleosomes 6 and 7) in the top panels are not picked up by our model. The annotation based on HMMD deviates significantly from the "hand picked" annotation.

5.2 Application to high resolution MNase-Chip and MNase-Seq data

Next, we will illustrate the applicability of our proposed NHSM in mapping nucleosome occupancy on the high resolution MNase-Chip (Lee et al.; 2007) and MNase-Seq data (Shivaswamy et al.; 2008). In both cases, we smoothed the data via Gaussian kernel smoothing described in Section 3 and defined $O_t = X_{t+4} - X_{t-5}$ $(k = 5)$. Details are given in Appendix A.2.4. We first demonstrate the utility of our proposed NHSM in annotating the *CHA1* and *HIS3* promoters in the 4 base pairs resolution MNase-Chip data of Lee et al.

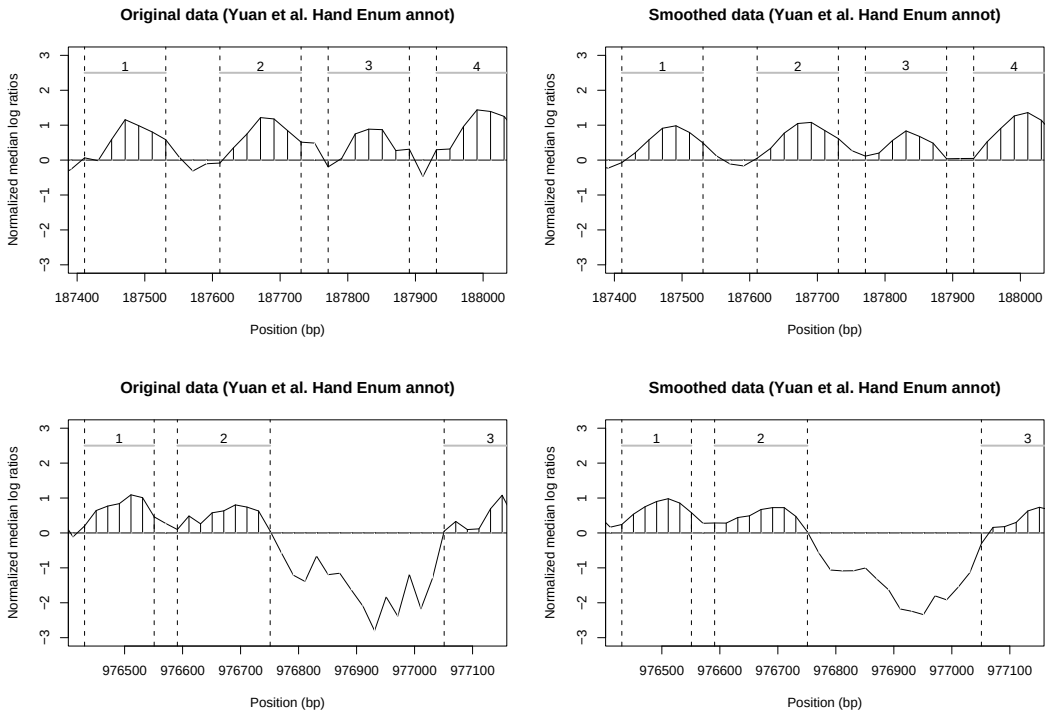


Figure 13: *Examples of "hand picked" annotations in Yuan et al. (2005).* Left panels are original data based on the "hand picked" annotations in Yuan et al. (2005) for two regions in chromosomes 5 and 7, respectively. Right panels are the smoothed data for similar regions. Although the "hand picked" nucleosomes are reliable, there are still some uncertainties in picking the boundaries of nucleosome-linker, for instance between nucleosomes 2 and 3 in the top panels and between nucleosomes 1 and 2 in the bottom panels.

(2007) (Figure 15). We observe that our proposed NHSM detects the low-signal nucleosomes between positions 17050 and 17200 in the *CHA1* promoter (labeled 1 in Figure 15) and between positions 722700 and 722850 in the *HIS3* promoter (labeled 2 in Figure 15), which were missed by the original SHMM annotation of Lee et al. (2007). These two low-signal nucleosomes were also identified by Shivaswamy et al. (2008) in their MNase-Seq data, indicating that they are not artifacts of hybridization.

Since our modeling framework utilizes first order differences which capture the "bump" shape of a nucleosome and not the observed log base 2 ratios in the emission distribution, it can be applied to first order (lagged) differences on tag counts/reads in MNase-Seq data. In Shivaswamy et al. (2008), 514803

DOI: 10.2202/1544-6115.1454

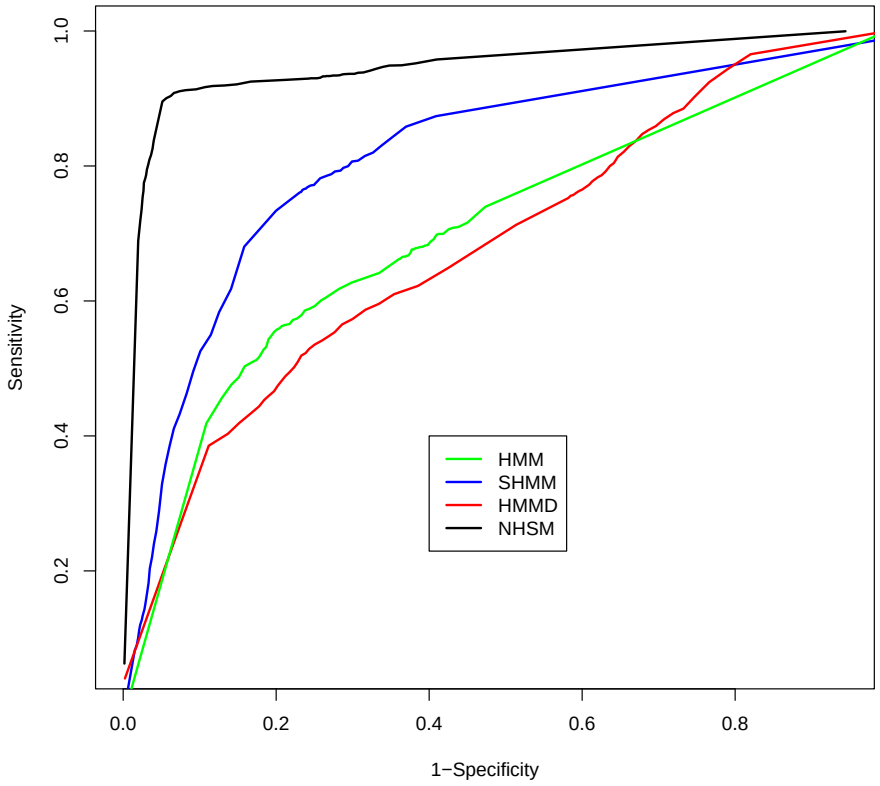


Figure 14: *Receiver operating characteristic (ROC) curve.* Comparison of various methods on MNase-chip data from Yuan et al. (2005) using the set of "hand picked" annotated low-signal nucleosomes as the true positive set.

uniquely aligned reads were generated for the normal cells via the sequencing technology. We considered the following strategy for mapping nucleosome positions on MNase-Seq data. Since each of the 27 base pairs Solexa sequencing read corresponds to a mono-nucleosome of size 150-200 base pairs, we first extended these reads to 150 base pairs according to the sequence orientation for both the plus and minus strands. The total reads for each genomic position is then taken to be the sum of all extended reads at the position, as shown in Figure 16. Therefore, the total reads at every 50 base pairs on the genome is analogous to the observed log base 2 ratios in MNase-chip data of 50 base pairs resolution.

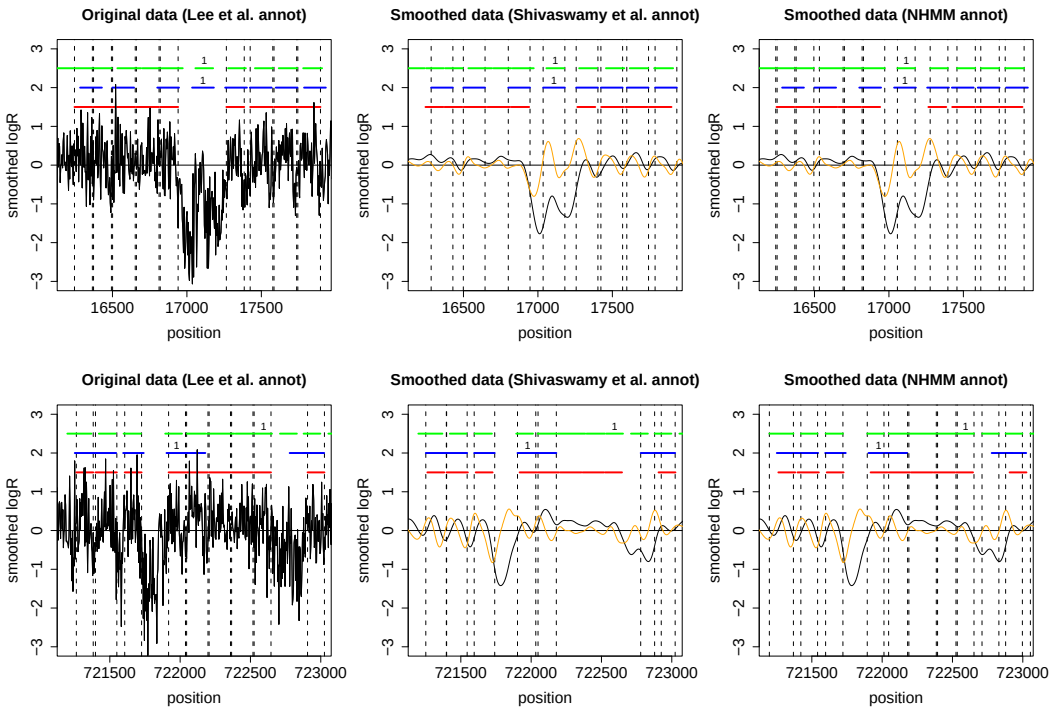


Figure 15: *Nucleosome occupancy at CHA1 (top row) and HIS3 (bottom row) promoter.* Red horizontal lines at y =1.5 are the nucleosome annotation from Lee et al. (2007). Blue horizontal lines at y =2 are the nucleosome annotation from Shivaswamy et al. (2008). Green horizontal lines at y =2.5 are the nucleosome annotation from NHSM. Vertical dotted lines in the left, middle, and right columns are boundaries separating nucleosome-linker states from Lee et al. (2007), Shivaswamy et al. (2008), and NHSM respectively, as given in the header. Orange lines are the computed O_t's for each mid-probe.

We also illustrate the applicability of our proposed NHSM in annotating these two promoter regions in the MNase-Seq data of Shivaswamy et al. (2008) using a 4 base pairs resolution to facilitate direct comparison against the 4 base pairs MNase-Chip data of Lee et al. (2007). Of particular interest is the ability of NHSM to decode three low-signal nucleosomes between positions 722200 and 722700 in the *HIS3* promoter (labeled 3, 4 and 5 in Figure 17), which were missed by the original annotation of Shivaswamy et al. (2008) as shown in Figure 17. The NHSM annotation for these three nucleosomes is consistent with the annotation from Lee et al. (2007). Both Figures 15 and 17 also illustrate that our proposed NHSM results in the most consistent, i.e.,

DOI: 10.2202/1544-6115.1454

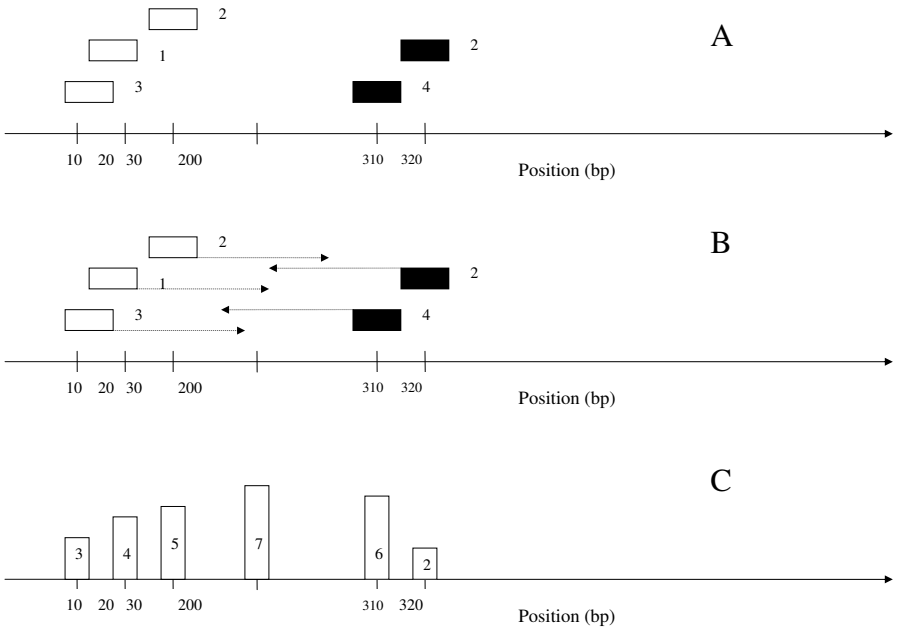


Figure 16: *Illustration of obtaining reads for each genomic position in ChIP-Seq data.* White rectangles are reads mapped to the plus strand and the black rectangles are reads mapped to the minus strand. Panel B shows the extended reads (150 base pairs). Panel C shows the total read for each genomic position.

greatest overlaps in annotations of nucleosome positions between MNase-Chip and MNase-Seq data. This is desirable given that both data sets measure the same nucleosome occupancy in yeast *S. cerevisiae*.

6 Discussion

The ability to map nucleosome positions accurately is crucial for investigating changes in nucleosome occupancies and their relationship to gene regulation since losses/gains in occupancy usually occur at one or two nucleosomes as illustrated in Shivaswamy et al. (2008). We introduced a non-homogeneous hidden-state model (NHSM) that automatically maps nucleosome positions based on either high-throughput tiling array or sequencing data and is computationally efficient. The modeling framework utilizes first order (lagged)

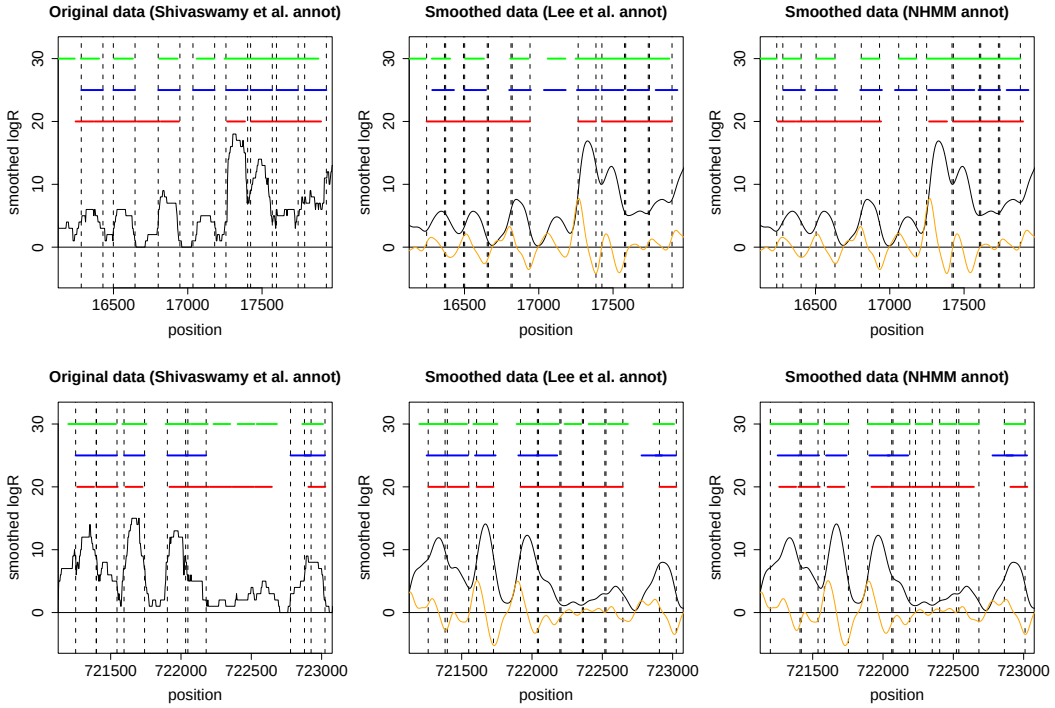


Figure 17: *Nucleosome occupancy at CHA1 (top row) and HIS3 (bottom row) promoter.* Red horizontal lines at y =20 are the nucleosome annotation from Lee et al. (2007). Blue horizontal lines at y =25 are the nucleosome annotation from Shivaswamy et al. (2008). Green horizontal lines at y =30 are the nucleosome annotation from NHSM. Vertical dotted lines in the left, middle, and right columns are boundaries separating nucleosome-linker states from Shivaswamy et al. (2008), Lee et al. (2007), and NHSM respectively, as given in the header. Orange lines are the computed O_t's for each mid-probe.

differences which capture the "bump" shape that characterize a nucleosome and enable accurate mapping of nucleosome-linker boundaries. The NHSM bypasses the need for further local detrending which misses low-signal nucleosomes (Figure 11). We also demonstrated the pitfalls of detrending the data with a simple method of comparing peak and trough within a window size covering a nucleosome (HMMD). Such a detrending introduced higher noise levels to the data in both the simulations and a case study of yeast nucleosome occupancy data. Modeling the emission distribution on first order differences allows our method to be applicable to both the MNase-Chip and MNase-Seq data, since the defining characteristic of a nucleosome in both cases is the "bump" shape. By allowing a duration distribution, NHSM is able to capture the fuzzy nucleosomes, which have heterogenic and dynamic positions.

DOI: 10.2202/1544-6115.1454

The only preprocessing step required before applying our proposed NHSM in detecting nucleosome positions is data smoothing. We have illustrated in the case studies that simple smoothing such as moving average in a window size of 3 is generally sufficient for lower resolution tiling arrays, e.g, the MNase-Chip data of Yuan et al. (2005). For high resolution tiling arrays and sequencing data which have lower signal-to-noise ratio, we proposed a Gaussian kernel smoothing with a bandwidth chosen based on the size of nucleosomal DNA, and demonstrated that this smoothing is able to denoise and enhance the "bump" shape of a nucleosome. Recently, Yassour et al. (2008) introduced a method for improving the resolution of the nucleosome positions in low resolution tiling arrays with overlapping probe design. In this paper, we aimed to develop a general method applicable to both low and high resolution nucleosome occupancy data by specifically capturing the defining characteristic, that is the "bump" shape of a nucleosome. However, the main idea of partitioning a probe into smaller fragments in Yassour et al. (2008) can be easily adapted into our framework by generating a pseudo MNase-Chip with higher resolution and thereby improving the resolution of nucleosome positions. We provide an illustration of this point in Appendix A.3.

The numerous examples and extensive simulations provided in this paper demonstrate that our proposed method is able to detect linker regions that are represented by only one/two probes, low-signal nucleosomes (Figures 10 and 11) and outperforms currently available methods. Although the underlying architecture of our NHSM is simple, it is effective in detecting nucleosome occupancies in both low and high resolution MNase-Chip and MNase-Seq data. Furthermore, in the datasets that we used for illustration, nucleosomal DNA was isolated by MNase but our approach is applicable to tiling arrays and sequencing technologies regardless of how the nucleosomes are isolated. We conclude by stressing that accurate annotation of nucleosome occupancy based on data from high-throughput experiments under various physiological conditions forms the basis of comparing different samples to elucidate the dynamics of nucleosome occupancy. Some examples of this line of work are by Shivaswamy et al. (2008) and Schones et al. (2008), and we anticipate that the number such examples will keep growing.

A Appendix

A.1 Choice of smoothing for nucleosome occupancy

We investigate various smoothing algorithms for denoising the observed log base 2 ratios in tiling arrays measuring nucleosome occupancy. We will illustrate the performance of the selected smoothing algorithms on the high resolution tiling arrays (4 base pairs resolution) from Lee et al. (2007). The original log base 2 ratios from one replicate/array for the *CHA1*, *HIS3* and *SAC7* promoters are shown in the top left panels of Figures 18, 19, and 20, respectively.

We have shown that data smoothing based on moving average in a window size $(2w + 1)$ of 3 probes works well in the lower resolution nucleosome data from Yuan et al. (2005). However, the data from Lee et al. (2007) has lower signal-to-noise ratio compared to Yuan et al. (2005). Therefore, one pass moving average using a window size of $2w + 1 = 3$ is not sufficient, as shown in top middle panels of Figures 18, 19, and 20. One possible solution is to use a larger w. Since a nucleosomal DNA is 146 base pairs long, a nucleosome occupied region will span approximately 32 probes. Hence, consider moving averages in a window size of 32 probes. As given in the top right panels of Figures 18, 19, and 20, using a larger w is able to increase the signal-to-noise ratio, although the denoised log base 2 ratios still exhibit some wiggly pattern.

Next, as an alternative for using a larger window size, we consider iterated moving averages using a window size of 3 probes. Let S and $S(Y)_t$ denote the smoother function and the resulting smoothed value at probe t, respectively. For example, $S(Y)_t = \sum_{j=t-w}^{t+w} Y_j/(2w+1)$ in the moving average approach. Let k denote the number of iterations and S_k be the resulting smoother function at the k-th iteration. Then,

$$
\begin{aligned}
S_1(Y)_t &= \frac{1}{3}(Y_{t-1} + Y_t + Y_{t+1}) \\
S_2(Y)_t &= \frac{1}{3^2}(Y_{t-2} + 2Y_{t-1} + 3Y_t + 2Y_{t+1} + Y_{t+2}) \\
S_3(Y)_t &= \frac{1}{3^3}(Y_{t-3} + 3Y_{t-2} + 6Y_{t-1} + 7Y_t + 6Y_{t+1} + 3Y_{t+2} + Y_{t+3}) \\
S_4(Y)_t &= \frac{1}{3^4}(Y_{t-4} + 4Y_{t-3} + 10Y_{t-2} + 16Y_{t-1} + 19Y_t + 16Y_{t+1} + 10Y_{t+2} \\
&\quad +4Y_{t+3} + Y_{t+4}) \\
&\vdots
\end{aligned}
$$

DOI: 10.2202/1544-6115.1454

The smoother function has the following general form:

$$S_k(Y)_t \;=\; \frac{1}{3^k}\sum_{j=t-k}^{t+k} c_j Y_j,$$

where $c_{t-j} = c_{t+j}$, $j = 1, ..., k$ (symmetry), $c_{t-k} \leq c_{t-k+1} \leq \ldots \leq c_{t-1}$ and $\sum_{j=t-k}^{t+k} c_j = 3^k$. The bottom left panels of Figures 18, 19, and 20 plot the resulting smoothed log base 2 ratios for the m-th iteration, where m is minimum k such that $|S_k(Y) - S_{k-1}(Y)|_{L_2} \leq \epsilon$, for $\epsilon = 10^{-6}$.

Next, we provide a simple analytical result for smoothing based on iterated moving averages. The collection of weights in front of the log base 2 ratios in S_k can be viewed as the resulting probability mass function from the sum of k independent and identically distributed discrete uniform random variables U_i taking values $\{-1, 0, 1\}$, denoted by f. In particular, f is related to weights in $S_k(Y)$ as follows:

$$f(\sum_{i=1}^{k} u_i) \;=\; \frac{c_j}{3^k} \text{ for } \sum_{i=1}^{k} u_i = j \in \{-k, -k+1, ..., k-1, k\}$$

We provide the form of f for $k = 2$ and $k = 3$:

$$f(u_1 + u_2) \;=\; \begin{cases} \frac{1}{9} \text{ for } u_1 + u_2 \in \{-2, 2\}, \\ \frac{2}{9} \text{ for } u_1 + u_2 \in \{-1, 1\}, \\ \frac{3}{9} \text{ for } u_1 + u_2 \in \{0\}. \end{cases}$$

$$f(u_1 + u_2 + u_3) \;=\; \begin{cases} \frac{1}{27} \text{ for } u_1 + u_2 + u_3 \in \{-3, 3\}, \\ \frac{3}{27} \text{ for } u_1 + u_2 + u_3 \in \{-2, 2\}, \\ \frac{6}{27} \text{ for } u_1 + u_2 + u_3 \in \{-1, 1\}, \\ \frac{7}{27} \text{ for } u_1 + u_2 + u_3 \in \{0\}. \end{cases}$$

Then, by the Central Limit Theorem,

$$\sum_{i=1}^{k} U_i/\sqrt{k} \rightarrow_D N(0, \sigma^2) \text{ as } k \rightarrow \infty,$$

where $\sigma^2 = (3^2 - 1)/12$. Therefore, for a sufficiently large fixed K,

$$f(\sum_{i=1}^{K} U_i = j) \approx \frac{\phi(j|0, K\sigma^2)}{\sum_{j=-K}^{K} \phi(j|0, K\sigma^2)},$$

i.e., a discretized Gaussian on integer support $\{-K, ..., K\}$, where $\phi(z|\mu, \sigma^2) = \exp\left[-(z-\mu)^2/2\sigma^2\right]/\sigma\sqrt{2\pi}$. In other words, the weights $c_j/3^K$, $j = 1, ..., K$, in $S_K(Y)_t$ are approximately $\phi(j|t, K\sigma^2)/\sum_{j=t-K}^{t+K}\phi(j|t, K\sigma^2)$. As the number of iterations (K) increases, $\phi(j|t, K\sigma^2) \to 0$. Therefore,

$$\begin{aligned} S_K(Y)_t &= \frac{1}{3^K}\sum_{j=t-K}^{t+K} c_j Y_j \\ &\approx \frac{\sum_{j=t-K}^{t+K}\phi(j|t, K\sigma^2)Y_j}{\sum_{j=t-K}^{t+K}\phi(j|t, K\sigma^2)}, \end{aligned}$$

and

$$\lim_{K\to\infty}\frac{\sum_{j=t-K}^{t+K}\phi(j|t, K\sigma^2)Y_j}{\sum_{j=t-K}^{t+K}\phi(j|t, K\sigma^2)} = \bar{Y}.$$

That is, the smoothed log base 2 ratios from iterated moving average become flatter and flatter approaching a constant $\bar{Y}$, and the choice of ϵ above is critical to avoid over-smoothing.

The analytical result above also shows that the iterated moving average is approximately a Gaussian kernel smoother for large k. In kernel smoothing, the tuning parameter is the bandwidth h. Large bandwidth implies more smoothing, and vice versa. For Gaussian kernel, h is also the standard deviation. That is,

$$X_t = \frac{\sum_{j=0}^{T}\phi\left(\frac{|G_t - G_j|}{h}\right)Y_j}{\sum_{j=0}^{T}\phi\left(\frac{|G_t - G_j|}{h}\right)},$$

where G_j is the genomic coordinate corresponding to Y_j in base pairs. We propose choosing h based on the size of the nucleosomal DNA. For a Gaussian distribution, 99% of the values are within $\pm 2.5\sigma$, where $\sigma = h$. We choose $h = 146/5$ so that the bandwidth spans the size of a nucleosome. The bottom right panels of Figures 18, 19, and 20 illustrate the resulting smoothed log base 2 ratios from Gaussian kernel smoothing with this choice of bandwidth. The Gaussian kernel smoothing is able to denoise the data and enhance the "bump" characteristics of a nucleosome, i.e., a series of decreasing positive slopes, followed by slopes of approximately zero in magnitude and then a series of increasing negative slopes.

DOI: 10.2202/1544-6115.1454

We also investigated another simple non-linear smoothing, i.e., iterated moving median (Tukey; 1977). This smoother is more robust and less sensitive to sudden jumps, and is usually recommended over moving average. Mallow (1979) proved that iterated moving median converges for odd-numbered spans (w). However, as shown in the bottom middle panels of Figures 18, 19, and 20, smoothing based on iterated moving median is less desirable since the "bump" shape of the nucleosomes is not well characterized compared to Gaussian kernel smoothing/iterated moving average. Another popular denoising algorithm is wavelet smoothing, which requires more tuning parameters (i.e., wavelet type, decomposition level and thresholds). The wavelet smoothing is utilized by Zhang et al. (2008) recently for detecting nucleosomes in ChIP-Seq. However, since we have illustrated that a Gaussian kernel smoothing is sufficient for denoising the log base 2 ratios from both the high resolution MNase-Chip and MNase-Seq data with a justified choice of bandwidth, we do not explore the more sophisticated wavelet smoothing.

A.2 Model fitting for NHSM with the expectation maximization algorithm

Let Q_t be the hidden state latent variable for mid-probe t and $\lambda = (\pi, A, B)$ denote the model parameters, where A is the transition probability matrix and B is the emission distribution. Also define $Z_t = O_t + Z_{t-1}$ for $t = 1, ..., T$ and Z_0 ($Z_0 = X_0$ if $O_t = X_t - X_{t-1}$ and $\tilde{Z}_0 = X_0 + ... + X_{2k-2}$ if $O_t = X_{t+k-1} - X_{t-k}$. See Appendix A.2.4), and let $O^{(T)} = (O_1, ..., O_T)$.

Two assumptions arising from Figure 4 are:

$$P(Q_{t+1}|Q^{(t)}, Z_0, O^{(t)}, \lambda) = P(Q_{t+1}|Q_t, Z_t, \lambda) \text{ for } t = 1, ..., T-1, \quad (3)$$

$$P(O_t|Q^{(t)}, Z_0, O^{(t-1)}, \lambda) = P(O_t|Q_t, \lambda) \text{ for } t = 1, ..., T. \quad (4)$$

For simplicity, we assume Z_0 is fixed. The complete data likelihood is given by:

$$\begin{aligned} P(O^{(T)}, Q^{(T)}|Z_0, \lambda) &= P(Q_1|Z_0, \lambda) \prod_{t=1}^{T} \left[P(O_t|Q^{(t)}, Z_0, O^{(t-1)}, \lambda)\right] \\ &\quad \prod_{t=1}^{T-1} \left[P(Q_{t+1}|Q^{(t)}, Z_0, O^{(t)}, \lambda)\right] \\ &= P(Q_1|Z_0, \lambda) \prod_{t=1}^{T} [P(O_t|Q_t, \lambda)] \prod_{t=1}^{T-1} [P(Q_{t+1}|Q_t, Z_t, \lambda)] \\ &= \pi_{Q_1} \prod_{t=1}^{T} b_{Q_t}(O_t) \prod_{t=1}^{T-1} a_{Q_t, Q_{t+1}}(Z_t). \end{aligned}$$

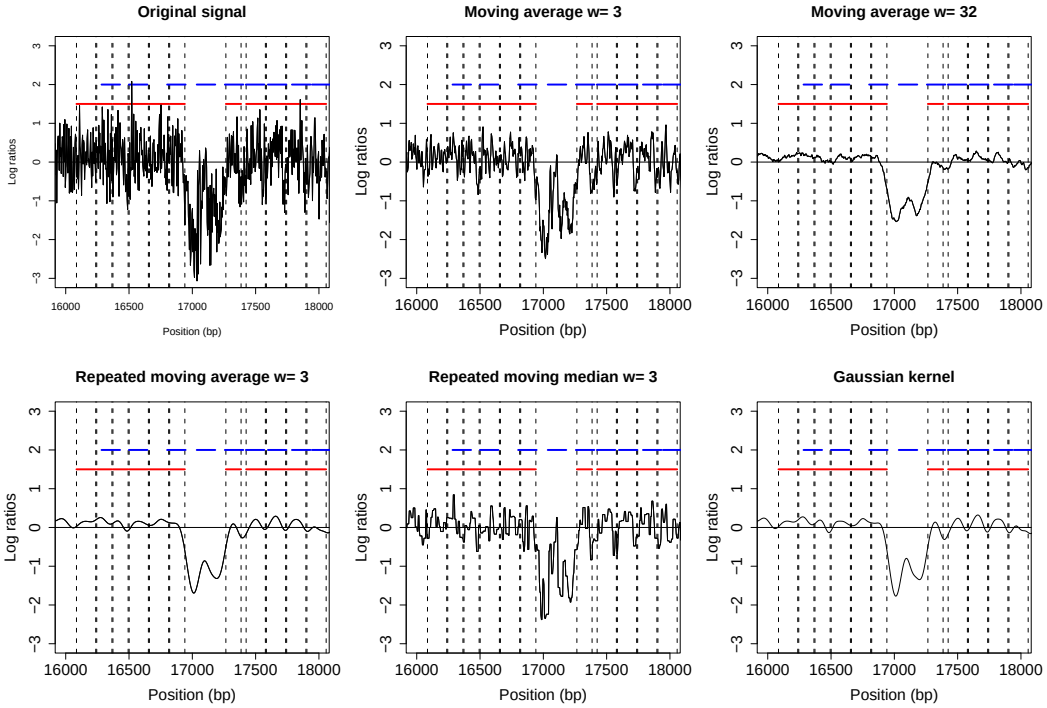


Figure 18: *Nucleosome occupancy in CHA1 promoter.* Different panels illustrate various smoothing algorithms. Vertical dotted lines are boundaries separating nucleosome-linker states from Lee et al. (2007). Red horizontal lines at y =1.5 are the nucleosome annotations from Lee et al. (2007). Blue horizontal lines at y =2 are the nucleosome annotations from Shivaswamy et al. (2008).

The first equality follows from repeated application of conditional probability. The second equality follows from the two assumptions above.

Assume that there are N hidden states. Then, the complete data log likelihood is given by:

$$\begin{aligned}\log P(O^{(T)}, Q^{(T)} \mid Z_0, \lambda) &= \log\left[\prod_{i=1}^{N} \pi_i^{I(Q_1=i)}\right]\left[\prod_{t=1}^{T}\prod_{i=1}^{N} b_i(O_t)^{I(Q_t=i)}\right] \\ &\quad \left[\prod_{t=1}^{T-1}\prod_{i=1}^{N}\prod_{j=1}^{N} a_{i,j}(Z_t)^{I(Q_t=i,Q_{t+1}=j)}\right].\end{aligned}$$

We assume that

$$\begin{aligned} b_i(O_t) &= N(\mu_i, \sigma_i^2), \\ a_{i,j}(Z_t) &= \frac{\exp(\gamma_{i,j} + \beta_j Z_t)}{\sum_{k=1}^{N} \exp(\gamma_{i,k} + \beta_k Z_t)}. \end{aligned}$$

DOI: 10.2202/1544-6115.1454

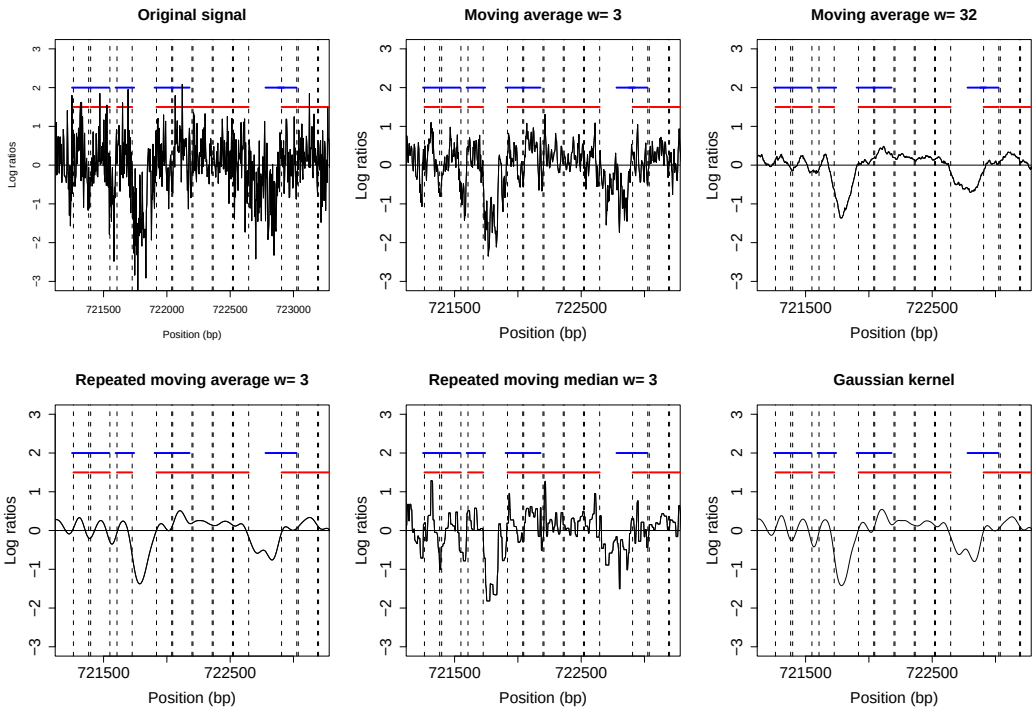


Figure 19: *Nucleosome occupancy in HIS3 promoter.* Different panels illustrate various smoothing algorithms. Vertical dotted lines are boundaries separating nucleosome-linker states from Lee et al. (2007). Red horizontal lines at y =1.5 are the nucleosome annotations from Lee et al. (2007). Blue horizontal lines at y =2 are the nucleosome annotations from Shivaswamy et al. (2008).

Expected complete log likelihood is given by

$$
\begin{aligned}
& E[\log P(O^{(T)}, Q^{(T)} \mid Z_0, \lambda)] \\
& = \sum_{i=1}^{N} P(Q_1 = i \mid O^{(T)}, Z_0, \lambda) \log \pi_i \\
& + \sum_{t=1}^{T} \sum_{i=1}^{N} P(Q_t = i \mid O^{(T)}, Z_0, \lambda) \log \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(O_t - \mu_i)^2}{2\sigma_i^2} \right) \right] \\
& + \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} P(Q_t = i, Q_{t+1} = j \mid O^{(T)}, Z_0, \lambda) \log a_{i,j}(Z_t).
\end{aligned}
$$

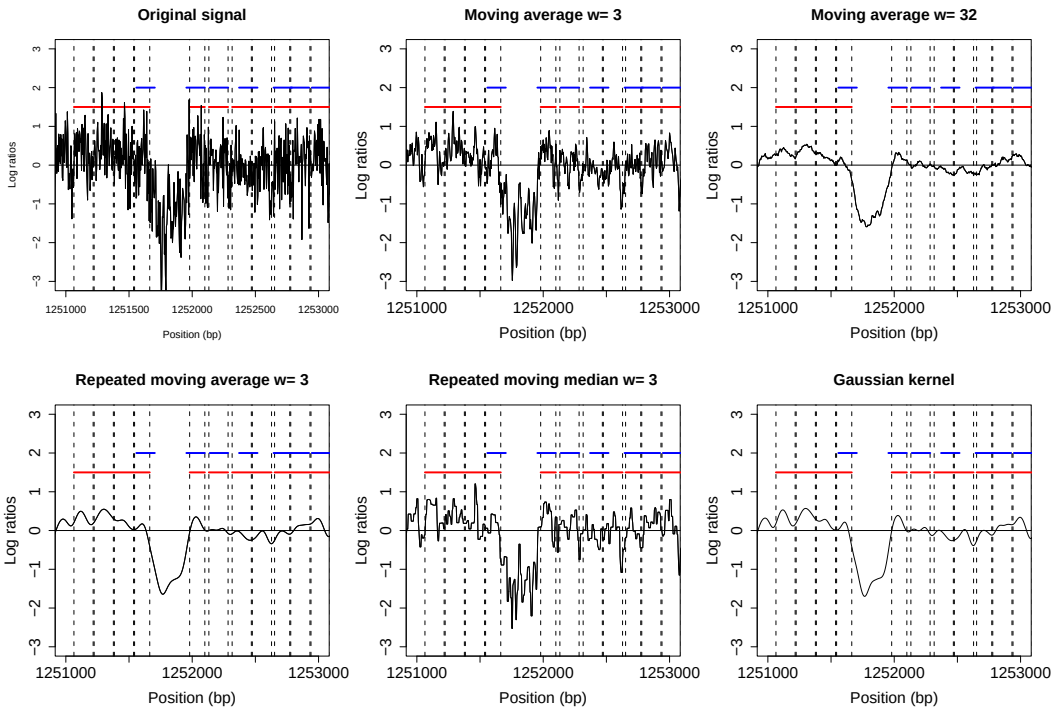


Figure 20: *Nucleosome occupancy in SAC7 promoter.* Different panels illustrate various smoothing algorithms. Vertical dotted lines are boundaries separating nucleosome-linker states from Lee et al. (2007). Red horizontal lines at y =1.5 are the nucleosome annotations from Lee et al. (2007). Blue horizontal lines at y =2 are the nucleosome annotations from Shivaswamy et al. (2008).

E-step:

Define two variables $\gamma_t(i)$ and $\xi_{t+1}(i,j)$:

$$\begin{aligned}
\gamma_t(i) &= P(Q_t = i \mid O^{(T)}, Z_0, \lambda) \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_T(i)} \\
&= \sum_{j=1}^{N} P(Q_t = i, Q_{t+1} = j \mid O^{(T)}, Z_0, \lambda),
\end{aligned}$$

DOI: 10.2202/1544-6115.1454

$$\begin{aligned}\xi_{t+1}(i,j) &= P(Q_t = i, Q_{t+1} = j \mid O^{(T)}, Z_t, \lambda) \\ &= \frac{\alpha_t(i)a_{i,j}(Z_t)b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{i,j}(Z_t)b_j(O_{t+1})\beta_{t+1}(j)} \\ &= \frac{\alpha_t(i)a_{i,j}(Z_t)b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\alpha_T(i)},\end{aligned}$$

where $\alpha_t(i) = P(O_1, ..., O_t, Q_t = i \mid Z_0, \lambda)$ and $\beta_t(i) = P(O_{t+1}, ..., O_T \mid Q_t = i, Z_0, \lambda)$.

M-step:

$$\max E[\log P(O^{(T)}, Q^{(T)} \mid Z_0, \lambda)]$$

$$\text{s.t} \quad \sum_{i=1}^{N}\pi_i = 1,$$

$$\sum_{j=1}^{N} a_{i,j}(Z_t) = 1, \quad t = 1, ..., T-1$$

yields

$$\pi_i = \gamma_1(i),$$

$$\begin{cases} \hat{\mu}_1 = \frac{\sum_{t=1}^{T}\gamma_t(B_N)O_t - \sum_{t=1}^{T}\gamma_t(B_L)O_t}{\sum_{t=1}^{T}\sum_{i\in\{B_N,B_L\}}\gamma_t(i)}, \\ \hat{\mu}_2 = \frac{\sum_{t=1}^{T}\sum_{i\in\{N_1,N_{2a},L_3\}}\gamma_t(i)O_t - \sum_{t=1}^{T}\sum_{i\in\{N_{2c},N_3,L_1\}}\gamma_t(i)O_t}{\sum_{t=1}^{T}\sum_{i\in\{N_1,N_{2a},N_{2c},N_3,L_1,L_3\}}\gamma_t(i)}, \quad \text{if } \hat{\mu}_1 \geq \hat{\mu}_2, \end{cases}$$

$$\hat{\mu}_1 = \hat{\mu}_2 = \frac{\sum_{t=1}^{T}\sum_{i\in\{B_N,N_1,N_{2a},L_3\}}\gamma_t(i)O_t - \sum_{t=1}^{T}\sum_{i\in\{B_L,N_{2c},N_3,L_1\}}\gamma_t(i)O_t}{\sum_{t=1}^{T}\sum_{i\in\{B_N,B_L,N_1,N_{2a},N_{2c},N_3,L_1,L_3\}}\gamma_t(i)},$$

if $\hat{\mu}_1 < \hat{\mu}_2$,

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T}\gamma_t(i)(O_t - \hat{\mu}_i)^2}{\sum_{t=1}^{T}\gamma_t(i)}.$$

The non-parametric transition probabilities for the case study are updated as follows:

$$\begin{aligned}\hat{a}_{i,j} &= \frac{\sum_{t=1}^{T-1}\xi_{t+1}(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \quad \text{for} \quad i \neq N_3, B_L, L_3, \\ \hat{a}^p_{N_3,B_L} &= \frac{\sum_{t=1}^{T-1}\xi_{t+1}(N_3,B_L)I(Z_t \geq 0)}{\sum_{t=1}^{T-1}\gamma_t(N_3)I(Z_t \geq 0)}, \\ \hat{a}^n_{B_L,B_N} &= \frac{\sum_{t=1}^{T-1}\xi_{t+1}(B_L,B_N)I(Z_t < 0)}{\sum_{t=1}^{T-1}\gamma_t(B_L)I(Z_t < 0)}, \\ \hat{a}^n_{L_3,B_N} &= \frac{\sum_{t=1}^{T-1}\xi_{t+1}(L_3,B_N)I(Z_t < 0)}{\sum_{t=1}^{T-1}\gamma_t(L_3)I(Z_t < 0)}.\end{aligned}$$

Note that the parameters in the logistic regression model cannot be solved analytically. These parameters can be optimized via a conjugate gradient algorithm. The details can be found in Robertson et al. (2004). The computation of $\alpha_t(i)$ and $\beta_t(i)$ is based on the well-known forward and backward procedures (Rabiner; 1989).

A.2.1 Forward procedure

Let $\alpha_t = P(O_1, O_2, ..., O_t, Q_t = i \mid Z_0, \lambda)$.

- Initialization:
  $\alpha_1(i) = P(O_1, Q_1 = i \mid Z_0, \lambda) = \pi_i b_i(O_1)$, for $1 \leq i \leq N$.
- Induction:
  $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{i,j}(Z_t)] b_j(O_{t+1})$, for $1 \leq t \leq T-1$, $1 \leq j \leq N$.
- Termination:
  $P(O^{(T)} \mid Z_0, \lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

A.2.2 Backward procedure

Let $\beta_t(i) = P(O_{t+1}, ..., O_T \mid Q_t = i, Z_0, \lambda)$.

- Initialization:
  $\beta_T(i) = 1$, for $1 \leq i \leq N$.
- Induction:
  $\beta_t(i) = \sum_{j=1}^{N} a_{i,j}(Z_t) b_j(O_{t+1}) \beta_{t+1}(j)$, for $t = T-1, T-2, ..., 1$, $1 \leq j \leq N$.
- Termination:
  $P(O^{(T)} \mid Z_0, \lambda) = \sum_{i=1}^{N} \beta_1(i) b_i(O_1) \pi_i$.

The optimal hidden state sequence is obtained via Viterbi algorithm (Rabiner; 1989).

A.2.3 Viterbi algorithm

Define $\delta_t(i) = \max_{Q_1,...,Q_{t-1}} P(Q_1, ..., Q_t = i, O_1, ..., O_t \mid Z_0, \lambda)$.

- Initialization:
  $\delta_1(i) = \pi_i b_i(O_1)$, for $1 \leq i \leq N$,
  $\psi_1(i) = 0$, for $1 \leq i \leq N$.

DOI: 10.2202/1544-6115.1454

- Recursion:
  $\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{i,j}(Z_t)]b_j(O_t)$, for $2 \le t \le T$, $1 \le j \le N$,
  $\psi_t(j) = \text{argmax}_{1\le i\le N}[\delta_{t-1}(i)a_{i,j}(Z_t)]$, for $2 \le t \le T$, $1 \le j \le N$.

- Termination:
  $P(O^{(T)} \mid Z_0, \lambda)^* = \max_{1\le i\le N}[\delta_T(i)]$,
  $Q_T^* = \text{argmax}_{1\le i\le N}[\delta_T(i)]$.

- Path backtracking:
  $Q_t^* = \psi_{t+1}(Q_{t+t}^*)$, for $t = T-1, T-2, ..., 1$.

A.2.4 Details of the NHSM for high resolution MNase-Chip and MNase-Seq data

For a general first order lagged differences $O_t = X_{t+k-1} - X_{t-k}$, O_t is defined for $t = k, ..., T-k+1$. We define $Z_t = O_t + Z_{t-1}$, and where $Z_t = X_{t+1-k} + ... + X_{t+k-1}$. To initialize the chain, let $Z_{k-1} = X_0 + ... + X_{2k-2}$. Let us define $\tilde{Z}_{t+1-k} = Z_t$ and $\tilde{O}_{t+1-k} = O_t$ so that the chain starts at $\tilde{O}_1$. Therefore, for the examples in Figure 15 and Figure 17:

- Initialize with $\tilde{Z}_0 = Z_4 = X_0 + ... + X_8$.
- Let $\tilde{O}_1 = O_5 = X_9 - X_0 \rightarrow \tilde{Z}_1 = Z_5 = X_1 + ... + X_9$.
- Let $\tilde{O}_2 = O_6 = X_{10} - X_1 \rightarrow \tilde{Z}_2 = Z_6 = X_2 + ... + X_{10}$ and so forth.

The transition probabilities for Figure 15 are:

$$
\begin{aligned}
a_{N_3,B_L}(\tilde{Z}_t) &= \begin{cases} 1, & \text{if } \tilde{Z}_t/(2k-1) < 0, \\ a^p_{N_3,B_L}, & \text{if } \tilde{Z}_t/(2k-1) \ge 0, \end{cases} \\
a_{B_L,B_N}(\tilde{Z}_t) &= \begin{cases} a^n_{B_L,B_N}, & \text{if } \tilde{Z}_t/(2k-1) < 0, \\ 1, & \text{if } \tilde{Z}_t/(2k-1) \ge 0, \end{cases} \\
a_{L_3,B_N}(\tilde{Z}_t) &= \begin{cases} a^n_{L_3,B_N}, & \text{if } \tilde{Z}_t/(2k-1) < 0, \\ 1, & \text{if } \tilde{Z}_t/(2k-1) \ge 0. \end{cases}
\end{aligned}
$$

The transition probabilities for Figure 17 are:

$$\begin{aligned}
a_{N_3,B_L}(\tilde{Z}_t) &= \begin{cases} 1, & \text{if } \tilde{Z}_t/(2k-1) < 5, \\ a^p_{N_3,B_L}, & \text{if } \tilde{Z}_t/(2k-1) \geq 5, \end{cases} \\
a_{B_L,B_N}(\tilde{Z}_t) &= \begin{cases} a^n_{B_L,B_N}, & \text{if } \tilde{Z}_t/(2k-1) < 5, \\ 1, & \text{if } \tilde{Z}_t/(2k-1) \geq 5, \end{cases} \\
a_{L_3,B_N}(\tilde{Z}_t) &= \begin{cases} a^n_{L_3,B_N}, & \text{if } \tilde{Z}_t/(2k-1) < 5, \\ 1, & \text{if } \tilde{Z}_t/(2k-1) \geq 5. \end{cases}
\end{aligned}$$

Here, the transition is based on $\tilde{Z}_t/(2k-1)$ which is the average log base 2 ratio in the window containing probes $t, t+1, ..., t+2k-2$. The mean tag count in Chromosome 3 is 5.2. Therefore, we choose 5 as the threshold for the transition probabilities.

A.3 Increasing resolution of MNase-Chip data

Recently, Yassour et al. (2008) introduced a method for mapping nucleosome positions, tailored for constant low resolution tiling arrays with overlapping probe design. Their method aimed to improve the resolution of the nucleosome positions by partitioning a probe into smaller fragments. This idea can be adapted into our model as follows: Let Y_{i-1}, Y_i and Y_{i+1} be the log base 2 ratios for 3 consecutive probes from a design similar to Yuan et al. (2005). Partition each Y_i into 5 segments $P^1_i, P^2_i, ..., P^5_i$ (Figure 21) as in Yassour et al. (2008). Since two probes overlap by 30 base pairs, P^3_{i-1} and P^1_i represent the same genomic region. Similarly, for each of (P^4_{i-1}, P^2_i), $(P^5_{i-1}, P^3_i, P^1_{i+1})$, (P^4_i, P^2_{i+1}) and (P^5_i, P^3_{i+1}), we have the same genomic region. We can create a pseudo MNase-Chip with 9 probes having log base 2 ratios R_i, where $R_1 = R_2 = Y_{i-1}$, $R_3 = R_4 = (Y_{i-1} + Y_i)/2$, $R_5 = (Y_{i-1} + Y_i + Y_{i+1})/3$, $R_6 = R_7 = (Y_i + Y_{i+1})/2$ and $R_8 = R_9 = Y_{i+1}$ as shown in Figure 21. After generating a pseudo MNase-Chip with higher resolution, our proposed NHSM can be readily used to detect nucleosome occupancy.

DOI: 10.2202/1544-6115.1454

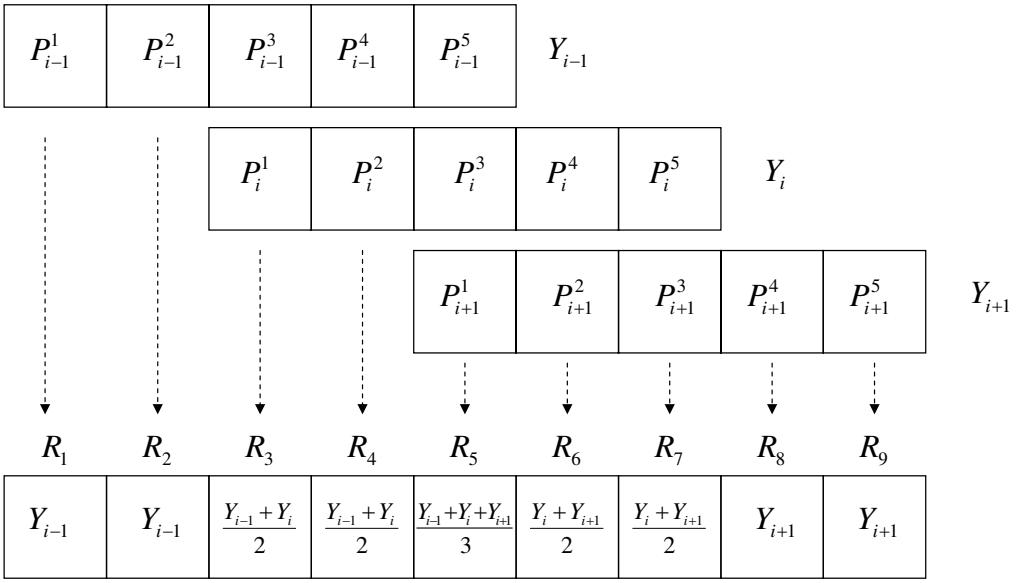


Figure 21: *Increasing resolution of tiling arrays via pseudo probes.* This is an illustration on how we can adapt the idea of Yassour et al. (2008) in creating a pseudo MNase-Chip data from constant low resolution tiling arrays with overlapping probe design. Y_{i-1}, Y_i and Y_{i+1} are the log base 2 ratios for 3 consecutive probes, whereas the P_i^j's are the resulting pseudo probes by partitioning each Y_i into 5 segments. R_i's are the resulting pseudo probes in the generated pseudo tiling array with higher resolution. The log base 2 ratios for this pseudo tiling array are obtained by averaging the original log base 2 ratios of the overlapping pseudo probes.

References

Albert, I., Mavrich, T., Tomsho, L., Qi, J., Zanton, S., Schuster, S. and Pugh, B. (2007). Translational and rotational settings of H2A.Z nucleosomes across the saccharomyces cerevisiae genome, *Nature* **446**: 572–576.

Bernstein, B., Liu, C., Humphrey, E., Perlstein, E. and Schreiber, S. (2004). Global nucleosome occupancy in yeast, *Genome Biology* **5**(62).

Chakravarthy, S., Park, Y., Chodaparambil, J., Edayathumangalam, R. and Luger, K. (2006). Structure and dynamic properties of nucleosome core particles, *FEBS Letters* **579**(4): 895–898.

Ercan, S., Carrozza, M. J. and Workman, J. L. (2004). Global nucleosome distribution and the regulation of transcription in yeast, *Genome Biology* **5**(10): doi:10.1186/gb–2004–5–10–243.

Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. and Fire, A. Z. (2006). Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin, *Genome Research* **16**: 1505–1516.

Kaplan, T., Liu, C., Erkmann, J., Holik, J., Grunstein, M., Kaufman, P., Friedman, N. and Rando, O. (2008). Cell cycle- and chaperone-mediated regulation of h3k56ac incorporation in yeast, *PLoS Genetics* .

Kornberg, R. and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome, *Cell* **98**: 285–294.

Lee, C., Shibata, Y., Rao, B., Strahl, B. and Lieb, J. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nature Genetics* .

Lee, W., Tillo, D., Bray, N., Morse, R., Davis, R., Hughes, T. and Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast, *Nature Genetics* .

Liu, C., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S., Friedman, N. and Rando, O. (2005). Single-nucleosome mapping of histone modifications in s.cerevisiae, *PLoS Biol* **3**(10): 1753–1769.

Mallow, C. (1979). *Some theoretical results on Tukeys 3R smoother*, Lecture Notes in Mathematics, Springer Berlin / Heidelberg.

Millar, C. and Grunstein, M. (2006). Genome-wide patterns of histone modifications in yeast, *Nature Reviews Molecular Cell Biology* **7**: 657–666.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286.

Robertson, A., Kirshner, S. and Smyth, P. (2004). Downscaling of daily rainfall occurrence over northeast brazil using a hidden markov model, *Journal of Climate* **17**(22): 4407–4424.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., Wei, G. and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome, *Cell* **132**(5): 887–898.

DOI: 10.2202/1544-6115.1454

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M. and Iyer, V. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation, *PLOS Biology* **6**(3): 618–630.

Shivaswamy, S. and Iyer, V. (2008). Stress-dependent dynamics of global chromatin remodeling in yeast: dual role for swi/snf in the heat shock stress response, *Molecular and celular biology* **28**(7): 2221–2234.

Tukey, J. (1977). *Exploratory data analysis*, Addison-Wesley.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J., Costa, G., McKernan, K., Sidow, A., Fire, A. and Johnson, S. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning, *Genome Research* **18**: 1051–1063.

Yassour, M., Kaplan, T., Jaimovich, A. and Friedman, N. (2008). Nucleosome positioning from tiling microarray data, *Bioinformatics* **24**: 139–146.

Yuan, G., Liu, Y., Dion, M., Slack, M., Wu, L., Altschuler, S. and Rando, O. (2005). Genome-scale idenfication of nucleosome positions in s.cerevisiae, *Science* **309**: 626–630.

Zhang, Y., Shin, H., Song, J., Lei, Y. and Liu, X. (2008). Identifying positioned nucleosomes with epigenetic marks in human from chip-seq, *BMC Genomics* **9**(1).

Zucchini, W., Raubenheimer, D. and MacDonald, I. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback, *Biometrics* **64**: 807–815.